Channel Coding and Source Coding With Increased Partial Side Information

Avihay Shirazi, Uria Basher and Haim Permuter

Abstract

Let $(S_{1,i}, S_{2,i}) \sim$ i.i.d $p(s_1, s_2)$, $i = 1, 2, \ldots$ be a memoryless, correlated partial side information sequence. In this work we study channel coding and source coding problems where the partial side information (S_1, S_2) is available at the encoder and the decoder, respectively, and, additionally, either the encoder's or the decoder's side information is increased by a limited-rate description of the other's partial side information. We derive six special cases of channel coding and source coding problems and we characterize the capacity and the rate-distortion functions for the different cases. We present a duality between the channel capacity and the rate-distortion cases we study. In order to find numerical solutions for our channel capacity and rate-distortion problems, we use the Blahut-Arimoto algorithm and convex optimization tools. As a byproduct of our work, we found a tight lower bound on the Wyner-Ziv solution by formulating its Lagrange dual as a geometric program. Previous results in the literature provide a geometric programming formulation that is only a lower bound, but not necessarily tight. Finally, we provide several examples corresponding to the channel capacity and the rate-distortion cases we presented.

Index Terms

Blahut-Arimoto algorithm, channel capacity, channel coding, convex optimization, duality, Gelfand-Pinsker channel coding, geometric programming, partial side information, rate-distortion, source coding, Wyner-Ziv source coding.

I. Introduction

In this paper we investigate point-to-point channel models and rate-distortion problem models where both users have different and correlated partial side information and where, in addition, a rate-limited description of one of the user's side information is delivered to the other user. We then show the duality between the channel models and the rate-distortion models we investigate. In the process of investigating the rate-distortion problems, we found a tight lower bound on the rate-distortion of the Wyner-Ziv [1] problem. We show here that it is possible to write the Lagrange dual of the Wyner-Ziv rate-distortion function as a geometric program. Then, we show that the optimal solution of this geometric program is the correct solution of the Wyner-Ziv problem.

For the convenience of the reader, we refer to the state information as the side information, to the partial side information that is available to the encoder as the encoder's side information (ESI) and to the partial side information that is available to the decoder as the decoder's side information (DSI). To the rate-limited description of the other

Avihay Shirazi, Uria Basher and Haim Permuter are with the Department of Electrical and Computer Engineering at the Ben Gurion University of the Negev, Beer Sheva, Israel. Emails: avihays@bgu.ac.il, basher@bgu.ac.il, haimp@bgu.ac.il The material in this paper was presented in part at the Allerton Conference on Communication, Control, and Computing, September 2010

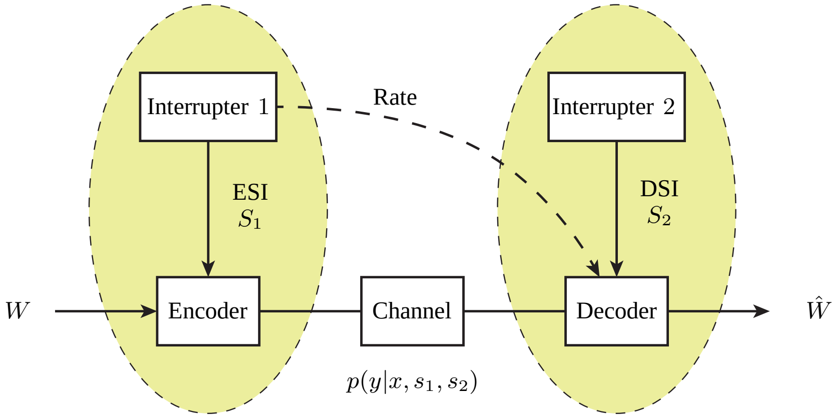

Fig. 1: *Increased partial side information example*. The encoder wants to send a message to the decoder over an interrupted channel in the presence of side information. The encoder is provided with the ESI and the decoder is provided with increased DSI. i.e., the decoder is informed with a rate-limited description of the ESI in addition to the DSI.

user's side information we refer as the increase in the side information. For example, if the decoder is informed with its DSI and, in addition, with a rate-limited description of the ESI, then we would say that the decoder is informed with *increased* DSI.

To make the motivation for this paper clear, let us look at a simple example, as depicted in Figure 1. Two remote users, User 1 - the encoder and User 2 - the decoder, want to communicate between them over a channel that is being interrupted by two interrupters, Interrupter 1 and Interrupter 2. We allow the interruptions S_1 and S_2 generated by the interrupters to be correlated, i.e., $(S_1, S_2) \sim p(s_1, s_2)$. Assume that Interrupter 1 is located in close proximity to User 1 and can fully describe its future interruption, S_1, to User 1 and that Interrupter 2 is located in close proximity to User 2 and can also fully describe its future interruption, S_2, to user 2. In addition, assume that Interrupter 1 can increase the side information of User 2 with rate-limited information about its interruption. In these circumstances, we pose the question; what is the capacity of the channel between User 1 and User 2? We extensively discuss the answer to this question in the forthcoming sections.

A. Channel capacity in the presence of state information

The three problems of channel capacity in the presence of state information that we adress in this paper are presented in Figure 2. We make the assumption that the encoder is informed with partial state information, the ESI (S_1), and the decoder is informed with different, but correlated, partial state information, which is the DSI (S_2). The channel capacity problem cases are:

- Case 1: The decoder is provided with increased DSI; i.e., in addition to the DSI, the decoder is also informed with a rate-limited description of the ESI.
- Case 2: The encoder is informed with increased ESI.
- Case 2_C: Similar to Case 2, with the exception that the ESI is known to the encoder in a causal manner. Notice that the rate-limited description of the DSI is still known to the encoder noncausally.

We will subsequently provide the capacity of Case 1 and Case 2_C and caracterize the lower and the upper bounds on Case 2, which differ only by a Markon relation. The results for the first case under discussion, Case 1, can be concluded from Steinberg's problem [2]. In [2], Steinberg introduced and solved the case in which the encoder is fully informed with the ESI and the decoder is informed with a rate-limited description of the ESI. Therefore, the innovation in Case 1 is that the decoder is also informed with the DSI. The solution for this problem can be derived by considering the DSI to be a part of the channel's output in Steinberg's solution. In the proof of the converse in his paper, Steinberg uses a new technique that involves using the Csiszár sum twice in order to get to a single-letter bound on the rate. We shall use this technique to present a duality in the converse of the Gelfand-Pinsker [3] and the Wyner-Ziv [1] problems, which, by themselves, constitute the basis for most of the results in this paper. In [1], Wyner and Ziv present the rate-distortion function for data compression problems with side information at the decoder. We make use of their coding scheme in the achievability proof of the lower bound of Case 2 for describing the ESI with a limited rate at the decoder. In [3], Gelfand and Pinsker present the capacity for a channel with noncausal CSI at the encoder. We use their coding scheme in the achievability proof of Case1 and the lower bound of Case 2 for transmitting information over a channel where the ESI is the state information at the encoder. Therefore, we combine in our problems the Gelfand-Pinsker and the Wyner-Ziv problems. Another related paper is [4], in which Shannon presented the capacity of a channel with causal CSI at the transmitter. We make use of Shannon's result in the achievability proof of Case 2_C for communicating over a channel with causal ESI at the encoder. We also use Shannon's strategies [4], for developing an iterative algorithm to calculate the capacity of the cases we present in this paper.

Some related papers that can be found in the literature are mentioned herein. Heegard and El Gamal [5] presented a model of a state-dependent channel, where the transmitter is informed with the CSI at a rate limited to R_e and the receiver is informed with the CSI at a rate limited to R_d. This result relates to Case 1, Case 2 and Case 2_C since we consider the rate-limited description of the ESI or the DSI as side information known at both the encoder and the decoder. Cover and Chiang [6] extended the Gelfand-Pinsker problem and the Wyner-Ziv problem to the case where both the encoder and the decoder are provided with different, but correlated, partial side information. They also showed a duality between the two cases, which is a topic that will be discussed later in this paper. Rozenzweig, Steinberg and Shamai [7] and Cemal and Steinberg [8] studied channels with partial state information at the transmitter. A detailed subject review on channel coding with state information was given by Keshet, Steinberg and Merhav in [9].

In addition to these three cases, we also present a more general case, where the encoder is informed with increased ESI and the decoder is informed with increased DSI; i.e., there is a rate-limited description of the ESI at the decoder and there is a rate-limited description of the DSI at the encoder. We provide an achievability scheme that bounds the capacity for this case from below, however, this bound does not coincide with the capacity and, therefore, this problem remains open.

B. *Rate-distortion with side information*

In this paper we adress three problems of rate-distortion with side information, as presented in Figure 3. In common with the channel capacity problems, we assume that the encoder is informed with the ESI (S_1) and the decoder is informed with the DSI (S_2), where the source, X, the ESI and the DSI are correlated. The rate-distortion problem cases we investigate in this paper are:

- Case 1: The decoder is provided with increased DSI.
- Case 1_C: Similar to Case 1, with the exception that the ESI is known to the encoder in a causal manner. The rate-limited description of the ESI is still known to the decoder noncausally.
- Case 2: The encoder is informed with increased ESI.

Case 2 is a special case of Kaspi's [10] two-way source coding for $K = 1$. In [10], Kaspi introduced a model of multistage communication between two users, where each user may transmit up to K messages to the other user, dependent on the source and the previous received messages. For Case 2, we can consider sending the rate-limited description of the DSI as the first transmission and then, sending a function of the source, the ESI and the rate-limited description of the DSI as the second transmission. This fits into Kaspi's problem for $K = 1$ and thus Kaspi's theorem also applies to Case 2. Kaspi's problem was later extended by Permuter, Steinberg and Weissman [11] to the case where a common rate-limited side information message is being conveyed to both users. Another strongly related paper is Wyner and Ziv's paper [1]. In the achievability of Case 1 we use the Wyner-Ziv coding scheme twice; once for describing the ESI at the decoder where the DSI is the side information and once for the main source and the ESI where the DSI is the side information. The rate-limited description of the ESI is the side information provided to both the encoder and the decoder. In [6] there is an extension to the Wyner-Ziv problem to the case where both the encoder and the decoder are provided with correlated partial side information. Weissman and El Gamal [12, Section 2] and Weissman and Merhav [13] presented source coding with causal side information at the decoder, which relates to Case 1_C.

As with the channel capacity, we present a bound on the general case of rate-distortion with two-sided increased partial side information. In this problem setup the encoder is informed with a rate-limited description of the DSI in addition to the ESI and the decoder is informed with a rate-limited description of the ESI in addition to the DSI. We present an achievability scheme that bounds the optimal rate from above, however, this bound does not coincide with the optimal rate and, therefore, this problem remains open.

C. *Duality*

Within the scope of this work we point out a duality relation between the channel capacity and the rate-distortion cases we discuss. The operational duality between channel coding and source coding was first mentioned by Shannon [14]. In [15], Pradhan, Chou and Ramchandran studied the functional duality between some cases of channel coding and source coding, including the duality between the Gelfand-Pinsker problem and the Wyner-Ziv problem. This duality was also described by Cover and Chiang in [6], where they provided a transformation that makes duality between channel coding and source coding with two-sided state information apparent. Zamir, Shamai and Erez [16] and Su, Eggers and Girod [17] utilized the duality between channel coding and source coding with side information to develop coding schemes for the dual problems.

In our paper we show that the channel capacity cases and the rate-distortion cases we discuss are operational duals in a way that strongly relates to the Wyner-Ziv and Gelfand-Pinsker duality. We also provide a transformation scheme that shows this duality in a clear way. Moreover, we show a duality relation between Kaspi's problem and Steinberg's [2] problem by showing a duality relation between Case 2 source coding and Case 1 channel coding. Also, we show duality in the converse parts of the Gelfand-Pinsker and the Wyner-Ziv problems. We show that both converse parts can be proven in a perfectly dual way by using the Csiszár sum twice.

D. Computational algorithms

Calculating channel capacity and rate-distortion problems, in general, and the Gelfand-Pinsker and the Wyner-Ziv problems, in particular, is not straightforward. Blahut [18] and Arimoto [19] suggested an iterative algorithm (to be referred to as the B-A algorithm) for numerically computing the channel capacity and the rate-distortion problems. Willems [20] and Dupuis, Yu and Willems [21] presented iterative algorithms based on the B-A algorithm for computing the Gelfand-Pinsker and the Wyner-Ziv functions. We use principles from Willems' algorithms to develop an algorithm to numerically calculate the capacity for the cases we presented. More B-A based iterative algorithms for computing channel capacity and rate-distortion with side information can be found in [22] and in [23]. A B-A based algorithm for maximizing the directed-information can be found in [24].

Another approach for solving the Wyner-Ziv rate-distortion problem is the geometric programming approach. This approach was presented by Chiang and Boyd in their paper [25], in which they described methods, based on convex optimization and geometric programming, to calculate the channel capacity of the Gelfand-Pinsker channel and to calculate a lower bound on the rate-distortion of the Wyner-Ziv problem. Chiang and Boyd considered the Lagrange-dual of the Wyner-Ziv problem and they formulated a geometric program that constitutes a lower bound on the rate-distortion. However, their lower bound is not tight because they implicitly used the assumption that the derivative of the Lagrangian is zero for each value of the side information individually, while the original expression is only restricted to zero when averaging over the side information. During our present work, we found a tight lower bound on the rate-distortion of the Wyner-Ziv problem. The tight bound is obtained by considering a primal variable in the dual problem. A similar trick has been used recently by Naiss and Permuter [26] for transforming the rate-distortion with feed-forward problem into a geometric program.

E. Organization of the paper and main contributions

To summarize, the main contributions of this paper are 1) we give single-letter characterizations of the capacity and the rate-distortion functions of new channel and source coding problems with increased partial side information, 2) we show a duality relationship between the channel capacity cases and the rate-distortion cases that we discuss, 3) we provide a tight lower bound on the Wyner-Ziv solution using convex optimization and geometric programming tools, 4) we provide a B-A based algorithm to solve the channel capacity problems we describe, 5) we show a duality between the Gelfand-Pinsker capacity converse and the Wyner-Ziv rate-distortion converse.

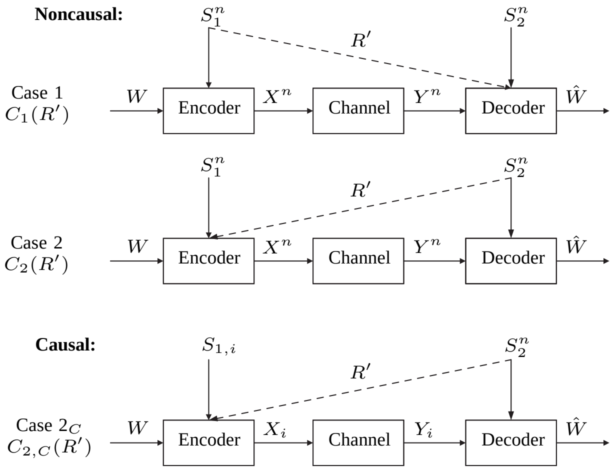

Fig. 2: Channel coding with state information. Case 1: Rate-limited ESI at the decoder. Case 2: Rate-limited DSI at the encoder. Case 2_C: Causal ESI and rate-limited DSI at the encoder.

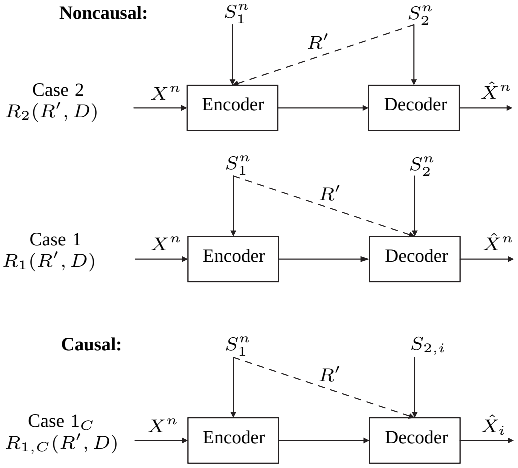

Fig. 3: Source coding with side information. Case 2: Rate-limited DSI at the encoder. Case 1: Rate-limited ESI at the decoder. Case 1_C: Causal DSI and rate-limited ESI at the decoder. The cases are presented in this order to allow each source coding case to be paralel to the dual channel coding case.

The reminder of this paper is organized as follows. In Section II we introduce some notations for this paper and provide the settings of three channel coding and three source coding cases with increased partial side information. In Section III we present the main results for coding with increased partial side information; we provide the capacity and the rate-distortion for the cases we introduced in Section II and we point out the duality between the cases we examined. Section IV contains the main results for the geometric programming; we formulate a geometric program that is a tight lower bound on the Wyner-Ziv solution. Section V contains illuminating examples for the cases discussed in the paper. In Section VI we describe the B-A based algorithm we used in order to solve the capacity examples. We conclude the paper in Section VII and we highlight two open problems; channel capacity and rate-distortion with two-sided rate-limited partial side information. Appendix A contains the duality derivation for the converse proofs of the Gelfand-Pinsker and the Wyner-Ziv problems and Appendices B through F contain the proofs for our theorems and lemmas.

II. Problem Setting and Definitions

In this section we describe and formally define three cases of channel coding problems and three cases of source coding problems. All six cases are presented in Figures 2 and 3.

Notations. We use subscripts and superscripts to denote vectors in the following ways: $x^j = (x_1, \ldots, x_j)$ and $x_i^j = (x_i, \ldots, x_j)$ for $i \leq j$. Moreover, we use the lower case x to denote sample value, the upper case X to denote a random variable, the calligraphic letter $\mathcal{X}$ to denote the alphabet of X, $|\mathcal{X}|$ to denote the cardinality of the alphabet of X and $p(x)$ to denote the probability $\Pr\{X = x\}$. We use the notation $\mathcal{T}_\epsilon^{(n)}(X)$ to denote the strongly typical set of the random variable X, as defined in [27, Chapter 11].

A. Definitions and problem formulation - channel coding with state information

Definition 1. A *discrete channel* is defined by the set $\{\mathcal{X}, \mathcal{S}_1, \mathcal{S}_2, p(s_1, s_2), p(y|x, s_1, s_2), \mathcal{Y}\}$. The channel's input sequence, $\{X_i \in \mathcal{X}, i = 1, 2, \dots\}$, the ESI sequence, $\{S_{1,i} \in \mathcal{S}_1, i = 1, 2, \dots\}$, the DSI sequence, $\{S_{2,i} \in \mathcal{S}_2, i = 1, 2, \dots\}$, and the channel's output sequence, $\{Y_i \in \mathcal{Y}, i = 1, 2, \dots\}$, are discrete random variables drawn from the finite alphabets $\mathcal{X}, \mathcal{S}_1, \mathcal{S}_2, \mathcal{Y}$, respectively. Denote the message and the message space as $W \in \{1, 2, \dots, 2^{nR}\}$ and let $\hat{W}$ be the reconstruction of the message W. The random variables $(S_{1,i}, S_{2,i})$ are i.i.d. $\sim p(s_1, s_2)$ and the channel is memoryless, i.e., at time i, the output, Y_i, has a conditional distribution of

$$p(y_i|x^i, s_1^i, s_2^i, y^{i-1}) = p(y_i|x_i, s_{1,i}, s_{2,i}). \tag{1}$$

In the remainder of the paper, unless specifically mentioned otherwise, we refer to the ESI and the DSI as if they are known to the encoder and the decoder, respectively, in a noncausal manner. Also, as noted before, we use the term *increased* side information to indicate that the user's side information also includes a rate-limited description of the other user's partial side information. For example, when the decoder is informed with the DSI and with a rate-limited description of the ESI we would say that the decoder is informed with *increased* DSI.

Problem Formulation. For the channel $p(y|x, s_1, s_2)$, consider the following channel coding problem cases:

- **Case 1**: The encoder is informed with ESI and the decoder is informed with increased DSI.
- **Case 2**: The encoder is informed with increased ESI and the decoder is informed with DSI.
- **Case 2$_C$**: The encoder is informed with increased causal ESI (S_1^i at time i) and the decoder is informed with DSI. This case is the same as Case 2, except for the causal ESI.

All cases are presented in Figure 2.

Definition 2. A $(n, 2^{nR}, 2^{nR'_j})$ *code*, $\{j \in 1, 2\}$, for a channel with increased partial side information, as illustrated in Figure 2, consists of two encoders and one decoder. The encoders are f and f_v, where f is the encoder for the channel's input and f_v is the encoder for the side information, and the decoder is g, as described for each case:
Case 1: Two encoders

$$\begin{aligned} f_v : &\quad \mathcal{S}_1^n \mapsto \{1, 2, \dots, 2^{nR'_1}\}, \\ f : &\quad \{1, 2, \dots, 2^{nR}\} \times \mathcal{S}_1^n \times \{1, 2, \dots, 2^{nR'_1}\} \mapsto \mathcal{X}^n, \end{aligned}$$

and a decoder

$$g : \quad \mathcal{Y}^n \times \mathcal{S}_2^n \times \{1, 2, \dots, 2^{nR'_1}\} \mapsto \{1, 2, \dots, 2^{nR}\}. \tag{2}$$

Case 2: Two encoders

$$\begin{aligned} f_v : &\quad \mathcal{S}_2^n \mapsto \{1, 2, \dots, 2^{nR'_2}\}, \\ f : &\quad \{1, 2, \dots, 2^{nR}\} \times \mathcal{S}_1^n \times \{1, 2, \dots, 2^{nR'_2}\} \to \mathcal{X}^n, \end{aligned}$$

and a decoder

$$g: \quad \mathcal{Y}^n \times \mathcal{S}_2^n \times \{1,2,\dots,2^{nR'_2}\} \mapsto \{1,2,\dots,2^{nR}\}. \tag{3}$$

Case 2_C: Two encoders

$$\begin{aligned} f_v: &\quad \mathcal{S}_2^n \mapsto \{1,2,\dots,2^{nR'_2}\}, \\ f_i: &\quad \{1,2,\dots,2^{nR}\} \times \mathcal{S}_1^i \times \{1,2,\dots,2^{nR'_2}\} \mapsto \mathcal{X}_i, \end{aligned}$$

and a decoder

$$g: \quad \mathcal{Y}^n \times \mathcal{S}_2^n \times \{1,2,\dots,2^{nR'_2}\} \mapsto \{1,2,\dots,2^{nR}\}. \tag{4}$$

The *average probability of error*, $P_e^{(n)}$, for a $(2^{nR}, 2^{nR'_j}, n)$ code is defined as

$$P_e^{(n)} = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \Pr\left\{\hat{W} \neq W \;|W = w\right\}, \tag{5}$$

where the index W is chosen according to a uniform distribution over the set $\{1,2,\dots,2^{nR}\}$. A rate pair (R, R') is said to be *achievable* if there exists a sequence of $(2^{nR}, 2^{nR'}, n)$ codes such that the average probability of error $P_e^{(n)} \to 0$ as $n \to \infty$.

Definition 3. The *capacity* of the channel, $C(R')$, is the supremum of all R such that the rate pair (R, R') is achievable.

B. Definitions and problem formulation - source coding with side information

Throughout this article we use the common definitions of rate-distortion as presented in [27].

Definition 4. The source sequence $\{X_i \in \mathcal{X}, i = 1,2,\dots\}$, the ESI sequence $\{S_{1,i} \in \mathcal{S}_1, i = 1,2,\dots\}$ and the DSI sequence $\{S_{2,i} \in \mathcal{S}_2, i = 1,2,\dots\}$ are discrete random variables drawn from the finite alphabets $\mathcal{X}, \mathcal{S}_1$ and $\mathcal{S}_2$ respectively. The random variables $(X_i, S_{1,i}, S_{2,i})$ are i.i.d $\sim p(x, s_1, s_2)$. Let $\hat{\mathcal{X}}$ be the reconstruction alphabet and $d_x : \mathcal{X} \times \hat{\mathcal{X}} \mapsto [0, \infty)$ be the distortion measure. The distortion between sequences is defined in the usual way:

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i). \tag{6}$$

Problem Formulation. For the source, X, the ESI, S_1, and the DSI, S_2, consider the following source coding problem cases:

- **Case 1**: The encoder is informed with ESI and the decoder is informed with increased DSI.
- **Case 2**: The encoder is informed with increased ESI and the decoder is informed with DSI.
- **Case 1_C**: The encoder is informed with ESI and the decoder is informed with increased causal DSI (S_2^i at time i). This case is the same as Case 1, except for the causal DSI.

All cases are presented in Figure 3.

Definition 5. A $(n, 2^{nR}, 2^{nR'_j}, D)$ *code*, $\{j \in 1, 2\}$, for the source X with increased partial side information, as illustrated in Figure 3, consists of two encoders, one decoder and a distortion constraint. The encoders are f and f_v, where f is the encoder for the source and f_v is the encoder for the side information, and the decoder is g, as described for each case:

Case 1: Two encoders

$$\begin{aligned} f_v: &\quad \mathcal{S}_1^n \mapsto \{1, 2, \ldots, 2^{nR'_1}\}, \\ f: &\quad \mathcal{X}^n \times \mathcal{S}_1^n \times \{1, 2, \ldots, 2^{nR'_1}\} \mapsto \{1, 2, \ldots, 2^{nR}\}, \end{aligned}$$

and a decoder

$$g: \quad \{1, 2, \ldots, 2^{nR}\} \times \mathcal{S}_2^n \times \{1, 2, \ldots, 2^{nR'_1}\} \mapsto \hat{\mathcal{X}}^n. \tag{7}$$

Case 2: Two encoders

$$\begin{aligned} f_v: &\quad \mathcal{S}_2^n \mapsto \{1, 2, \ldots, 2^{nR'_2}\}, \\ f: &\quad \mathcal{X}^n \times \mathcal{S}_1^n \times \{1, 2, \ldots, 2^{nR'_2}\} \mapsto \{1, 2, \ldots, 2^{nR}\}, \end{aligned}$$

and a decoder

$$g: \quad \{1, 2, \ldots, 2^{nR}\} \times \mathcal{S}_2^n \times \{1, 2, \ldots, 2^{nR'_2}\} \mapsto \hat{\mathcal{X}}^n. \tag{8}$$

Case 1_C: Two encoders

$$\begin{aligned} f_v: &\quad \mathcal{S}_1^n \mapsto \{1, 2, \ldots, 2^{nR'_1}\}, \\ f: &\quad \mathcal{X}^n \times \mathcal{S}_1^n \times \{1, 2, \ldots, 2^{nR'_1}\} \mapsto \{1, 2, \ldots, 2^{nR}\}, \end{aligned}$$

and a decoder

$$g_i: \quad \{1, 2, \ldots, 2^{nR}\} \times \mathcal{S}_2^i \times \{1, 2, \ldots, 2^{nR'_1}\} \mapsto \hat{\mathcal{X}}_i. \tag{9}$$

The distortion constraint for all three cases is:

$$\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} d(X_i, \hat{X}_i)\Big] \leq D. \tag{10}$$

For a given distortion, D, and for any $\epsilon > 0$, the rate pair (R, R') is said to be *achievable* if there exists a $(n, 2^{nR}, 2^{nR'}, D + \epsilon)$ code for the rate-distortion problem.

Definition 6. For a given R' and distortion D, the *operational rate* $R^*(R', D)$ is the infimum of all R, such that the rate pair (R, R') is achievable.

III. Coding with Increased Partial Side Information - Main Results

In this section we present the main results of this paper. We will first present the results for the channel coding cases, then the main results for the source coding cases and, finally, we will present the duality between them.

A. *Channel coding with side information*

For a channel with two-sided state information as presented in Figure 2, where $(S_{1,i}, S_{2,i}) \sim p(s_1, s_2)$, the *capacity* is as follows

Theorem 1 (The capacity for the cases in Figure 2)**.** For the memoryless channel $p(y|x, s_1, s_2)$, where S_1 is the ESI and S_2 is the DSI and the side information $(S_{1,i}, S_{2,i}) \sim p(s_1, s_2)$, the channel capacity is

Case 1: The encoder is informed with ESI and the decoder is informed with increased DSI,

$$C_1^* = \max_{\substack{p(v_1|s_1)p(u|s_1,v_1)p(x|u,s_1,v_1) \\ \text{s.t.} \quad R' \geq I(V_1;S_1) - I(V_1;Y,S_2)}} I(U;Y,S_2|V_1) - I(U;S_1|V_1). \tag{11}$$

Case 2: The encoder is informed with increased ESI and the decoder is informed with DSI;
Lower bounded by

$$C_2^{lb*} = \max_{\substack{p(v_2|s_2)p(u|s_1,v_2)p(x|u,s_1,v_2) \\ \text{s.t.} \quad R' \geq I(V_2;S_2|S_1)}} I(U;Y,S_2|V_2) - I(U;S_1|V_2). \tag{12}$$

Upper bounded by

$$C_2^{ub1*} = \max_{\substack{p(v_2|s_1,s_2)p(u|s_1,v_2)p(x|u,s_1,v_2) \\ \text{s.t.} \quad R' \geq I(V_2;S_2|S_1)}} I(U;Y,S_2|V_2) - I(U;S_1|V_2) \tag{13}$$

and by

$$C_2^{ub2*} = \max_{\substack{p(v_2|s_2)p(u|s_1,s_2,v_2)p(x|u,s_1,v_2) \\ \text{s.t.} \quad R' \geq I(V_2;S_2|S_1)}} I(U;Y,S_2|V_2) - I(U;S_1|V_2). \tag{14}$$

Case 2$_C$: The encoder is informed with increased causal ESI (S_1^i at time i) and the decoder is informed with DSI,

$$C_{2C}^* = \max_{\substack{p(v_2|s_2)p(u|v_2)p(x|u,s_1,v_2) \\ R' \geq I(V_2;S_2)}} I(U;Y,S_2|V_2). \tag{15}$$

For case j, $j \in \{1,2\}$, some joint distribution, $p(s_1, s_2, v_j, u, x, y)$, and (U, V_j) being some auxiliary random variables with bounded cardinality.

Section B contains the proof.

Lemma 1. For all three channel coding cases described in this section and for $j \in \{1,2\}$, the following statements hold

(i) The function $C_j(R')$ is a concave function of R'.

(ii) It is enough to take X to be a deterministic function of (U, S_1, V_j) to evaluate C_j.

(iii) The auxiliary alphabets $\mathcal{U}$ and $\mathcal{V}_j$ satisfy

$$\begin{array}{lc}
\text{for Case 1:} & |\mathcal{V}_1| \leq |\mathcal{X}||\mathcal{S}_1||\mathcal{S}_2| + 1 \quad \text{and} \\
 & |\mathcal{U}| \leq |\mathcal{X}||\mathcal{S}_1||\mathcal{S}_2|\big(|\mathcal{X}||\mathcal{S}_1||\mathcal{S}_2| + 1\big), \\
\hline
\text{for Case 2:} & |\mathcal{V}_2| \leq |\mathcal{S}_1||\mathcal{S}_2| + 1 \quad \text{and} \\
 & |\mathcal{U}| \leq |\mathcal{X}||\mathcal{S}_1||\mathcal{S}_2|\big(|\mathcal{S}_1||\mathcal{S}_2| + 1\big), \\
\hline
\text{for Case } 2_C\text{:} & |\mathcal{V}_2| \leq |\mathcal{S}_2| + 1 \quad \text{and} \\
 & |\mathcal{U}| \leq |\mathcal{X}||\mathcal{S}_2|\big(|\mathcal{S}_2| + 1\big).
\end{array}$$

Appendix D contains the proof for the above lemma.

Remark: We assume that the lower bound of Case 2 is tight, namely, $C_2 = C_2^{lb}$. This claim is hard to corroborate; we have not, as yet, derived a converse proof that maintains both Markov relations $V_2 - S_2 - S_2$ and $U - (S_1, V_2) - S_2$ and that bounds any achievable rate from above simultaneously.

B. *Source coding with side information*

For the problem of source coding with side information as presented in Figure 3, the *rate-distortion* function is as follows:

Theorem 2 (The rate-distortion function for the cases in Figure 3)**.** For a bounded distortion measure $d(x, \hat{x})$, a source, X, and side information, S_1, S_2, where $(X_i, S_{1,i}, S_{2,i}) \sim p(x, s_1, s_2)$, the rate-distortion function is

Case 1: The encoder is informed with ESI and the decoder is informed with increased DSI,

$$R_1^*(D) = \min_{\substack{p(v_1|s_1)p(u|x,s_1,v_1)p(\hat{x}|u,s_2,v_1) \\ \text{s.t.} \quad R' \geq I(V_1;S_1|S_2)}} I(U; X, S_1|V_1) - I(U; S_2|V_1). \tag{16}$$

Case 1$_C$: The encoder is informed with ESI and the decoder is informed with increased causal DSI (S_2^i at time i),

$$R_{1C}^*(D) = \min_{\substack{p(v_1|s_1)p(u|x,s_1,v_1)p(\hat{x}|u,s_2,v_1) \\ \text{s.t.} \quad R' \geq I(V_1;S_1)}} I(U; X, S_1|V_1). \tag{17}$$

Case 2: The encoder is informed with increased ESI and the decoder is informed with DSI,

$$R_2^*(D) = \min_{\substack{p(v_2|s_2)p(u|x,s_1,v_2)p(\hat{x}|u,s_2,v_2) \\ \text{s.t.} \quad R' \geq I(V_2;S_2) - I(V_2;X,S_1)}} I(U; X, S_1|V_2) - I(U; S_2|V_2). \tag{18}$$

For case j, $j \in \{1, 2\}$, some joint distribution, $p(x, s_1, s_2, v_j, u, \hat{x})$, where $\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} d(X_i, \hat{X}_i)\Big] \leq D$ and (U, V_j) being some auxiliary random variables with bounded cardinality.

Section C contains the proof.

Lemma 2. For all cases of rate-distortion problems in this section and for $j \in \{1, 2\}$, the following statements hold.

(i) The function $R_j(R', D)$ is a convex function of R' and D.

(ii) It is enough to take $\hat{X}$ to be a deterministic function of (U, S_2, V_j) to evaluate R_j.

(iii) The auxiliary alphabets $\mathcal{U}$ and $\mathcal{V}_j$ satisfy

for Case 1:	$\lvert\mathcal{V}_1\rvert \leq \lvert\mathcal{S}_1\rvert\lvert\mathcal{S}_2\rvert + 1$ and
	$\lvert\mathcal{U}\rvert \leq \lvert\mathcal{X}\rvert\lvert\mathcal{S}_1\rvert\lvert\mathcal{S}_2\rvert\big(\lvert\mathcal{S}_1\rvert\lvert\mathcal{S}_2\rvert + 1\big),$
for Case 1_C:	$\lvert\mathcal{V}_1\rvert \leq \lvert\mathcal{S}_1\rvert + 1$ and
	$\lvert\mathcal{U}\rvert \leq \lvert\mathcal{X}\rvert\lvert\mathcal{S}_1\rvert\big(\lvert\mathcal{S}_1\rvert + 1\big),$
for Case 2:	$\lvert\mathcal{V}_2\rvert \leq \lvert\mathcal{X}\rvert\lvert\mathcal{S}_1\rvert\lvert\mathcal{S}_2\rvert + 1$ and
	$\lvert\mathcal{U}\rvert \leq \lvert\mathcal{X}\rvert\lvert\mathcal{S}_1\rvert\lvert\mathcal{S}_2\rvert\big(\lvert\mathcal{X}\rvert\lvert\mathcal{S}_1\rvert\lvert\mathcal{S}_2\rvert + 1\big).$

Appendix D contains the proof for the above lemma.

C. Main results - duality

We now investigate the duality between the channel coding and the source coding for the cases in Figures 2 and 3. The following transformation makes the duality between the channel coding cases 1, 2, 2_C and the source coding cases 2, 1, 1_C, respectively, evident. The left column corresponds to channel coding and the right column to source coding. For cases j and $\bar{j}$, where $j, \bar{j} \in \{1, 2\}$ and $\bar{j} \neq j$, consider the transformation:

$$\text{channel coding} \longleftrightarrow \text{source coding} \tag{19}$$

$$C \longleftrightarrow R(D) \tag{20}$$

$$\text{maximization} \longleftrightarrow \text{minimization} \tag{21}$$

$$C_j \longleftrightarrow R_{\bar{j}}(D) \tag{22}$$

$$X \longleftrightarrow \hat{X} \tag{23}$$

$$Y \longleftrightarrow X \tag{24}$$

$$S_j \longleftrightarrow S_{\bar{j}} \tag{25}$$

$$V_j \longleftrightarrow V_{\bar{j}} \tag{26}$$

$$U \longleftrightarrow U \tag{27}$$

$$R' \longleftrightarrow R'. \tag{28}$$

This transformation is an extension of the transformation provided in [6] and in [15]. Note that while the channel capacity formula in Case j and the rate-distortion function in Case $\bar{j}$ are dual to one another in the sense of maximization-minimization, the corresponding rates R' are not dual to each other in this sense; i.e., one would expect to see an opposite inequality ($\geq \leftrightarrow \leq$) for dual cases, where we have an inequality that is in the same direction ($\leq \leftrightarrow \leq$) in the R' formulas. The duality in the side information rates, R', is then in the sense that the arguments in the formulas for the dual R' are dual. This exception is due to the fact that while the Gelfand-Pinsker and the Wyner-Ziv problems for the main channel or the main rate-distortion problems are dual, the Wyner-Ziv problem for the side information stays the same; the only difference is the input and the output.

IV. Geometric Programming

In this section, we provide a method to evaluate the Wyner-Ziv rate, using the Lagrange dual function and geometric programming. Before presenting the main results on this subject, let us provide the definitions and notations that we will use throughout this section and throughout the proof of the forthcoming main results.

A. Definitions and preliminaries - convex optimization and Lagrange duality

Most of the notations and the definitions that we use in this section are taken from [28]. We denote the variable x with dimension greater than 1 as $\mathbf{x}$ and we use $\mathbf{x} \succeq 0$ to denote that $x_i \geq 0$ for all $i = 1, 2, \ldots, \dim(\mathbf{x})$.

Consider the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & f_0(\mathbf{x}) \\ \text{subject to} \quad & f_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \ldots, m, \\ & h_j(\mathbf{x}) = 0, \quad j = 1, 2, \ldots, p, \end{aligned} \tag{29}$$

with the variable $\mathbf{x} \in \mathbb{R}^n$. We refer to f_0 as the *objective function* of the optimization problem and to f_i and h_j as the *constraint functions*. We let $\mathcal{D}$ denote the domain of $\mathbf{x}$; this is the set of all points for which the objective and the constraint functions are defined. We denote the optimal minimizer of $f_0(\mathbf{x})$ in $\mathcal{D}$ as $\mathbf{x}^*$. If the objective function, $f_0(\mathbf{x})$, and the inequality constraint functions, $f_i(\mathbf{x}),\ i = 1, 2, \ldots, m$, are all convex in $\mathbf{x}$ and the equality constraint functions, $h_j(\mathbf{x}),\ j = 1, 2, \ldots, p$, are affine in $\mathbf{x}$, then the problem is said to be a *convex optimization problem*. The *Lagrangian* associated with problem (29) is

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{x}), \tag{30}$$

where $\mathbf{x} \in \mathcal{D},\ \boldsymbol{\lambda} \in \mathbb{R}^m$ and $\boldsymbol{\mu} \in \mathbb{R}^p$. The *Lagrange dual function*, as defined in [28, Capter 5.1.2], is

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\mathbf{x} \in \mathcal{D}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}). \tag{31}$$

Following from [28, Chapter 5.1.3], for any $\boldsymbol{\lambda}$ where $\lambda_i \geq 0$ for $i = 1, 2, \ldots, m$, the Lagrange dual function yields a lower bound on the optimal value, $f_0(\mathbf{x}^*)$. The *Lagrange dual problem* [28, Chapter 5.2] associated with (29) is

$$\begin{aligned} \text{maximize} \quad & g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \\ \text{subject to} \quad & \lambda_i \geq 0, \quad i = 1, 2, \ldots, m. \end{aligned} \tag{32}$$

In this context, we refer to the original problem (29) as the *primal problem*. The *strong duality* property is associated with the case where the solution for the dual problem and the solution for the primal problem coincide. Following from [28, Chapter 5.2.3], if the primal problem is convex and Slater's condition [28, Chapter 5.2.3] holds, then strong duality holds.

A special family of optimization problems that we are interested in is the family of geometric programs. This type of optimization problems is defined in [28, Chapter 4.5] and is summarized here. Define *monomial* as the function

$$f(\mathbf{x}) = c x_1^{a_1} x_2^{a_2} \dots x_n^{a_n}, \tag{33}$$

were $c > 0$ and $a_i \in \mathbb{R}$. A sum of monomials, i.e., a function of the form

$$f(\mathbf{x}) = \sum_{k=1}^{K} c_k x_1^{a_{1k}} x_2^{a_{2k}} \dots x_n^{a_{nk}}, \tag{34}$$

where $c_k > 0$, is called a *posynomial*. An optimization problem of the form

$$\begin{aligned} \text{minimize} \quad & f_0(\mathbf{x}) \\ \text{subject to} \quad & f_i(\mathbf{x}) \leq 1, \quad i = 1, 2, \dots, m, \\ & h_j(\mathbf{x}) = 1, \quad j = 1, 2, \dots, p, \end{aligned} \tag{35}$$

where $f_0, \dots, f_m$ are posynomials, $h_1, \dots, h_p$ are monomials and $\mathbf{x} \succeq 0$ is called a *geometric program*. Geometric programs, as mentioned in [28, Chapter 4.5], are not convex problems. However, these problems can be transformed into convex optimization problems by taking $\log(\cdot)$ on both the objective and the constraint functions.

B. Problem Setting and Main Results

Let us consider the classic Wyner-Ziv problem as illustrated in Figure 4. Assume correlated random variables $(X, S) \sim$ i.i.d. $p(x, s)$ with finite alphabets $\mathcal{X}, \mathcal{S}$, respectively. Let $\{(X_i, S_i)\}_{i=1}^n$ be a sequence of n independent drawings of (X, S). Let the sequence X^n be the source sequence and let S^n be the side information sequence available at the decoder. We wish to describe the source, X, at rate R bits per symbol and to reconstruct $\hat{X}$ at the decoder with a distortion smaller than or equal to D, i.e., when encoding X in blocks of length n, we desire that $\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n d(X_i, \hat{X}_i)\Big] \leq D$.

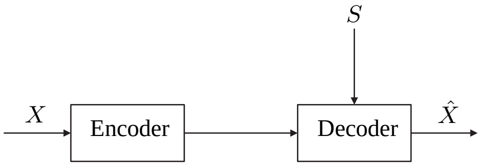

Fig. 4: *The Wyner-Ziv problem.*

The rate-distortion function with side information at the decoder [1] is

$$R(D) = \min_{p(u|x)p(\hat{x}|u,s)} I(U; X|S) \tag{36}$$

for some joint distribution $p(x, s, u, \hat{x})$ such that $\mathbb{E}\Big[d(X, \hat{X})\Big] \leq D$, i.e., $\sum_{x,s,u,\hat{x}} p(x, s)p(u|x)p(\hat{x}|u, s)d(x, \hat{x}) \leq D$. According to [20], we can write the expression of the rate-distortion function as

$$R(D) = \min_{q(t|x)} I(T; X|S) \tag{37}$$

for some joint distribution $p(x,s,t)=p(x,s)q(t|x)$, where $\mathcal{T}$ is the set of all mappings

$$t: \quad \mathcal{S}\mapsto\hat{\mathcal{X}}, \tag{38}$$

and the distortion constraint

$$\sum_{x,s,t} p(x,s)q(t|x)d\big(x,t(s)\big)\leq D \tag{39}$$

is maintained. We denote the set of $q(t|x)$'s for all $x\in\mathcal{X}$ and $t\in\mathcal{T}$ as $\mathbf{q}\in\mathbb{R}^{|\mathcal{T}||\mathcal{X}|}$ and we note that $I(T;X|S)$ is a convex function of $\mathbf{q}$ and that the rate-distortion function, $R(D)$, is its optimal value.

Combining (37) and (39), we get that the Wyner-Ziv problem is the following problem

$$\begin{array}{ll} \text{minimize} & \sum_{x,s,t} p(x,s)q(t|x)\log\frac{q(t|x)}{Q(t|s)} \\ \text{subject to} & \sum_t q(t|x)=1 \quad \forall x, \\ & \sum_{x,s,t} p(x,s)q(t|x)d\big(x,t(s)\big)\leq D, \\ & q(t|x)\geq 0 \quad \forall x,t, \end{array} \tag{40}$$

where the variables of the optimization are $\mathbf{q}$ and the constant parameters are the source distribution, $p(x,s)$, the distortion measure, $d\big(x,t(s)\big)$, and the distortion constraint, D, for all $x\in\mathcal{X}$, $s\in\mathcal{S}$ and $t\in\mathcal{T}$. The marginal distribution $Q(t|s)$ is defined by

$$Q(t|s)=\frac{\sum_x p(x,s)q(t|x)}{\sum_x p(x,s)}, \tag{41}$$

We define the set of $Q(t|s)$'s for all $s\in\mathcal{S}$ and $t\in\mathcal{T}$ as $\mathbf{Q}\in\mathbb{R}^{|\mathcal{T}||\mathcal{S}|}$.

The main result of this section is brought in the following theorem.

Theorem 3. The Lagrange dual of the Wyner-Ziv rate-distortion problem is the following geometric program (in convex form):

$$\begin{array}{ll} \text{maximize} & \sum_x p(x)\alpha_x-\gamma D \\ \text{subject to} & \alpha_x+\sum_s p(s|x)\Big[\log p(x|s)-\gamma d\big(x,t(s)\big)-y_{x,s,t}\Big]\leq 0 \quad \forall x,t, \\ & \log\left(\sum_x \exp\left\{y_{x,s,t}\right\}\right)\leq 0 \quad \forall s,t, \\ & \gamma\geq 0, \end{array} \tag{42}$$

where the optimization variables are $\boldsymbol{\alpha}\in\mathbb{R}^{|\mathcal{X}|},\gamma\in\mathbb{R}_+$ and $\mathbf{y}\in\mathbb{R}^{|\mathcal{X}||\mathcal{S}||\mathcal{T}|}$, and the constant parameters are the source distribution $p(x,s)$, the distortion measure $d\big(x,t(s)\big)$ and the distortion constraint, D. Furthermore, if Slater's condition [28, Chapter 5.2.3] holds, then strong duality holds and the solution for the optimization problem in (42) is a tight lower bound on the Wyner-Ziv solution, (40), and $R(D)$ is its optimal value.

Proof: The proof for Theorem 3 is given in Appendix E.

V. EXAMPLES

In this section we provide examples for Case 2 of the channel coding theorem and for Case 1 of the source coding theorem. The numerical iterative algorithm, which we used to numerically calculate the lower bound, C_2^{lb}, is provided in the next section.

Example 1 (*Case 2 channel coding for a binary channel)***.** Consider the binary channel illustrated in Figure 5. The alphabet of the input, the output and the two states is binary $\mathcal{X} = \mathcal{Y} = \mathcal{S}_1 = \mathcal{S}_2 = \{0, 1\}$ with $(S_1, S_2) \sim \mathbf{P_{S_1S_2}}$ being a joint PMF matrix. The channel is dependent on the states S_1 and S_2, where the encoder is fully informed with S_1 and with S_2 with a rate limited to R' and the decoder is fully informed with S_2. The dependence of the channel on the states is illustrated in Figure 5. If $(S_1 = 1, S_2 = 0)$ then the channel is the *Z channel* with transition probability ϵ, if $(S_1 = 1, S_2 = 1)$ then the channel has no error, if $(S_1 = 0, S_2 = 0)$ then the channel is the *X-channel* and if $(S_1 = 0, S_2 = 1)$ then the channel is the *S-channel* with transition probability of ϵ. The side information's joint pmf is

$$\mathbf{P_{S_1S_2}} = \begin{pmatrix} 0.1 & 0.4 \\ 0.4 & 0.1 \end{pmatrix}.$$

The expressions for the lower bound on the capacity $C_2^{lb}(R')$ and for R' are brought in Case 2 of Theorem 1.

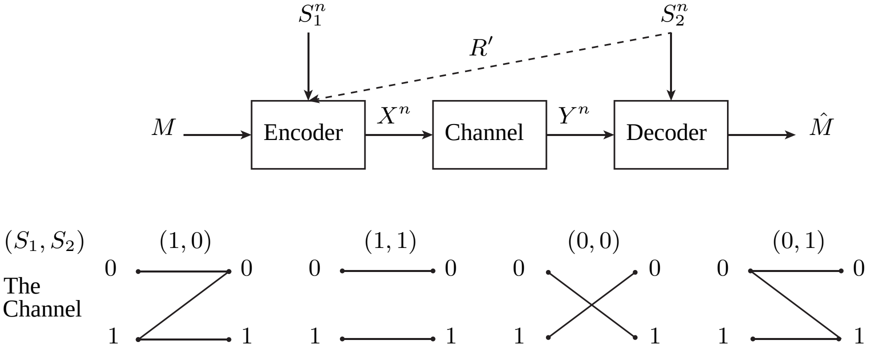

Fig. 5: Example 1 *Channel coding Case 2* - channel topology.

In Figure 6 we provide the graph from of the computation of the lower bound on the capacity for the binary channel we are testing. In the graph, we present the lower bound, $C_2^{lb}(R')$, as a function of R'. We also provide the Cover & Chiang [6] capacity (where $R' = 0$) and the Gelfand & Pinsker [3] capacity (where $R' = 0$ and the decoder is not informed with S_2).

Discussion:

1) The algorithm that we used to calculate $C_2^{lb}(R')$ and R' combines a grid-search and a Blahut-Arimoto-like

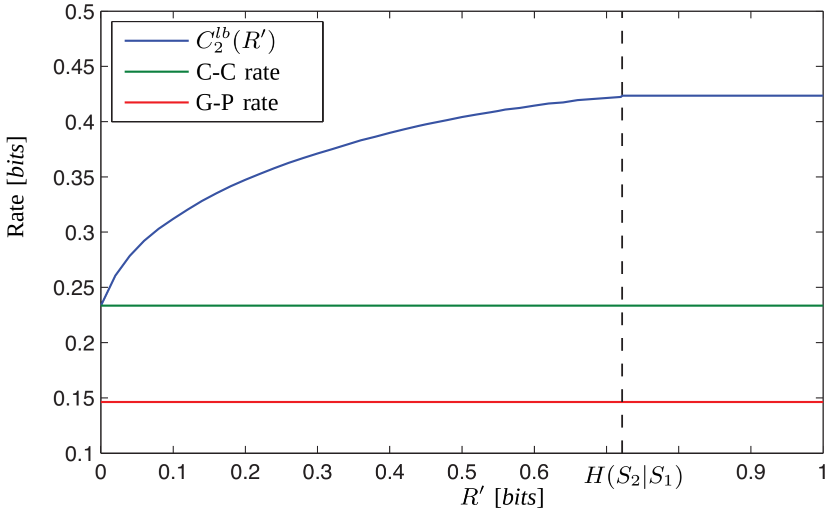

Fig. 6: Example 1. *Channel coding Case 2* for the channel depicted in Figure 5, where the side information is distributed $S_1 \sim \text{Bernoulli}(0.5)$, and $\Pr\{S_2 \neq S_1\} = 0.8$. $C_2^{lb}(R')$ is the lower bound on the capacity of this channel, *C-C rate* is the Cover-Chiang rate ($R' = 0$) and *G-P rate* is the Gelfand-Pinsker rate ($R' = 0$ and the decoder has no side information available at all). Notice that at the encoder the maximal uncertainty about S_2 is $H(S_2|S_1) = 0.7219$ *bit*. Therefore, for any $R' \geq 0.7219$ C_2^{lb} reaches its maximal value.

algorithms. We first construct a grid of probabilities of the random variable V_2 given S_2, namely, $w(v_2|s_2)$. Then, for every probability $w(v_2|s_2)$ such that $I(V_2; S_2|S_1)$ is close enough to R' we calculate the maximum of $I(U; Y, S_2|V_2) - I(U; S_1|V_2)$ using the iterative algorithm described in the next section. We then choose the maximum over those maximums and declare it to be C_2^{lb}. By taking a fine grid of the probabilities $w(v_2|s_2)$ the operation's result can be arbitrarily close to C_2^{lb}.

2) For a given joint PMF matrix $\mathbf{P_{S_1S_2}}$, we can see that $C_2^{lb}(R')$ is non-decreasing in R'. Furthermore, since the expression $I(V_2; S_2|S_1)$ is bounded by $R_{\max} = \max_{p(v_2|s_2)} I(V_2; S_2|S_1) = H(S_2|S_1)$, allowing R' to be greater than $R_{\max}$ cannot improve C_2^{lb} any more. i.e., $C_2^{lb}(R' = R_{\max}) = C_2^{lb}(R' > R_{\max})$. Therefore, it is enough to allow $R' = R_{\max}$ to achieve C_2^{lb}, as if the encoder is fully informed with S_2.

3) Although C_2^{lb} is a lower bound on the capacity, it can be significantly greater than the Cover-Chiang and the Gelfand-Pinsker rates for some channel models, as can be seen in this example. Moreover, we can actually state that C_2^{lb} is always greater than or equal to the Gelfand-Pinsker and the Cover-Chiang rates. This is due to the fact that when $R' = 0$, C_2^{lb} coincides with the Cover-Chiang rate, which, in its turn, is always greater than or equal to the Gelfand-Pinsker rate; since C_2^{lb} is also non-decreasing in R', it is obvious that our assertion holds.

Example 2 (*Source coding Case 1 for a binary-symmetric source and Hamming distortion)***.** Consider the source $X = S_1 \oplus S_2$, where $S_1, S_2 \sim$ i.i.d. $\text{Bernoulli}(0.5)$, and consider the problem setting depicted in Case 1 of the source coding problems. It is sufficient for the decoder to reconstruct S_1 with distortion $\mathbb{E}\big[d(S_1, \hat{S}_1)\big] \leq D$ in order to reconstruct X with the same distortion. Furthermore, the two rate-distortion problem settings illustrated in Figure 7 are equivalent.

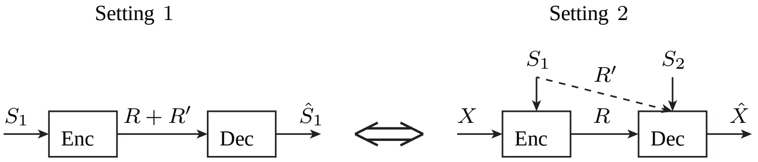

Fig. 7: The equivalent rate-distortion problem for Case 1 for the source $X = S_1 \oplus S_2$ where $S_1, S_2 \sim$ i.i.d. Bernoulli(0.5).

For every achievable rate in Setting 1, $\mathbb{E}\Big[d(S_1, \hat{S}_1)\Big] \leq D$. Denote $\hat{X} \triangleq \hat{S}_1 \oplus S_2$, then, $d(S_1, \hat{S}_1) = S_1 \oplus \hat{S}_1 = (S_1 \oplus S_2) \oplus (\hat{S}_1 \oplus S_2) = X \oplus \hat{X} = d(X, \hat{X})$ and, therefore, $\mathbb{E}\Big[d(S_1, \hat{S}_1)\Big] \leq D$ in Setting 1 $\Rightarrow$ $\mathbb{E}\Big[d(X, \hat{X})\Big] \leq D$ in Setting 2. In the same way, for Setting 2, denote $\hat{S}_1 \triangleq \hat{X} \oplus S_2$. Then, $d(X, \hat{X}) = X \oplus \hat{X} = S_1 \oplus \hat{S}_1$ and, therefore, $\mathbb{E}\Big[d(X, \hat{X})\Big] \leq D$ in Setting 2 $\Rightarrow$ $\mathbb{E}\Big[d(S_1, \hat{S}_1)\Big] \leq D$ in Setting 1. Hence, we can conclude that the two settings are equivalent and, for any given $0 \leq D$ and $0 \leq R'$, the rate-distortion function is

$$R(D) = \begin{cases} 1 - H(D) - R' & 1 - H(D) - R' \geq 0 \\ 0 & 1 - H(D) - R' < 0 \end{cases} . \tag{43}$$

In Figure 8 we present the plot resulting for this example. It is easy to verify that the Wyner & Ziv rate and the Cover & Chiang rate for this setting are $R_{WZ}(D) = R_{CC}(D) = \max\big\{1 - H(D), 0\big\}$.

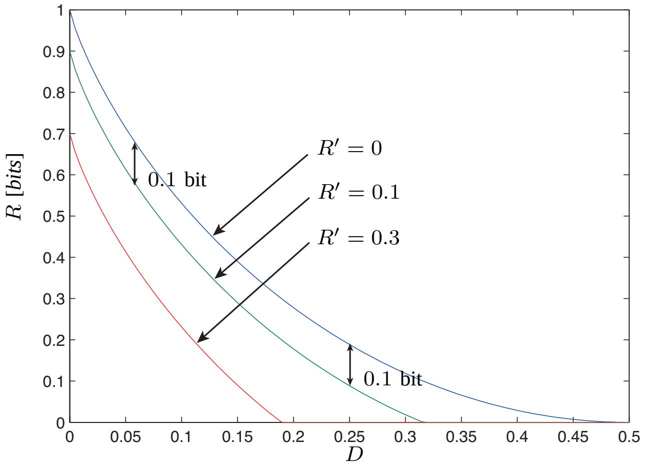

Fig. 8: Example 2. *Source coding Case 1 for binary-symmetric source and Hamming distortion*. The source is given by $X = S_1 \oplus S_2$, where $S_1, S_2 \sim$ Bernoulli(0.5). The graph shows the rate-distortion function for different values of R'.

Example 3 (*Geometric programming and the Wyner-Ziv problem)***.** Consider the traditional Wyner-Ziv [1] problem where the source, X, and the side information, S, are distributed according to $X \sim \text{Bernoulli}(0.5)$ and $\Pr\{S \neq X\} = 0.3$. We calculated the rate-distortion function, $R(D) = \min_{p(u|x)p(\hat{x}|u,s)} I(U;X|S)$ s.t. $\mathbb{E}\Big[d(X, \hat{X}) \leq D\Big]$, by using three different methods: first by using [1, Theorem II], second by using [25, Proposition 3] and third by using the geometric programming solution we introduced in Theorem 3. The plot resulting from this computation is brought in Figure 9.

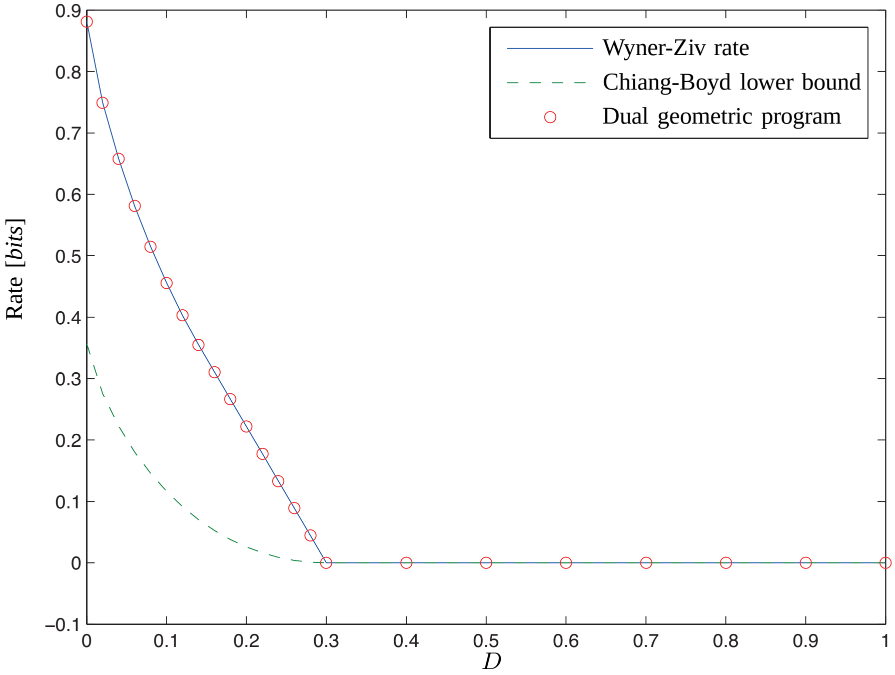

Fig. 9: Example 3. *Geometric programming and Wyner-Ziv.* The source and the side information distribute $X \sim \mathrm{Bernoulli}(0.5)$ and $\Pr\{S \neq X\} = 0.3$.

It can be seen in the figure that the geometric program, which was calculated according to Theorem 3, is tight to the Wyner-Ziv rate.

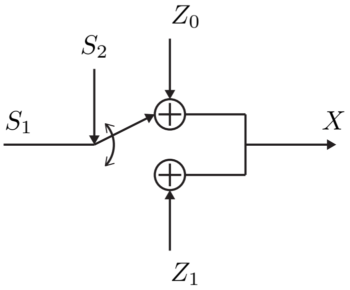

Fig. 10: Example 4. *Source coding Case 1* with binary symmetric source generation, as given in (44)

Example 4 (*Geometric programming and source coding Case 1)***.** Again, consider a rate-distortion problem as outlined in Case 1 with a binary-symmetric source and Hamming distortion. The source, X, is the output of the system illustrated in Figure 10, $S_1, S_2 \sim$ i.i.d. $\mathrm{Bernoulli}(0.5)$, S_2 is controlling a switch, $Z_0 \sim$ Bernoulli (0.3) and $Z_1 \sim$ Bernoulli (0.001). The output of this system can be expressed as

$$X = \begin{cases} S_1 \oplus Z_0, & S_2 = 0 \\ S_1 \oplus Z_1, & S_2 = 1 \end{cases}. \tag{44}$$

This source coding problem was introduced by Cheng, Stankovic and Xiong [22] for the case where the users are not allowed to share with each other their partial side information ($R' = 0$). The rate-distortion expression for this problem is $R_1(D) = \min I(U;X,S_1|V_1) - I(U;S_2|V_1)$, where the minimization is over all $p(v_1|s_1)p(u|x,s_1,v_1)p(\hat{x}|u,s_2,v_1)$ s.t. $R' \geq I(V_1;S_1|S_2)$ and that $\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} d(X_i,\hat{X}_i)\Big] \leq D$. We solve this example by using the geometric programming expression we developed in Theorem 3. The algorithm we developed in order to solve this problem uses some of the main principles we used in the algorithm that we developed for Example 1 (Algorithm 1) and that is detailed in Section VI. For this reason, we now bring a summary of the algorithm for this example.

First, as claimed in Section IV, it is possible to write the expression for the rate-distortion as $R(D) = \min I(T;X,S_1|V_1) - I(T;S_2|V_1)$ where the minimization is over all $w(v_1|s_1)q(t|x,s_1,v_1)$ s.t. $R' \geq I(V_1;S_1|S_2)$ and that $\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} d(X_i,T(S_2,V_1))\Big] \leq D$. The variable T is the mapping $T : \mathcal{S}_2 \times \mathcal{V}_1 \to \hat{\mathcal{X}}$. It can be verified that for every fixed probability, $w(v_1|s_1)$, the function $I(T;X,S_1|V_1) - I(T;S_2|V_1)$ is a convex function of $q(t|x,s_1,v_1)$. Now, we construct a fine grid of probabilities $w(v_1|s_1)$, and we keep those $w(v_1|s_1)$ for which $R' \geq I(V_1;S_1|S_2) \geq R' - \epsilon$ in the array $\mathcal{W}^*$. At this point, for every $w(v_1|s_1) \in \mathcal{W}^*$ that we kept, we let $R_w(D)$ be the solution for the following geometric program

$$\begin{array}{ll} \text{maximize} & \sum_{x,s_1,v_1} \alpha_{x,s_1,v_1} p(x,s_1,v_1) - \gamma D \\ \text{subject to} & \alpha_{x,s_1,v_1} + \sum_{s_2} p(s_2|x,s_1)\Big[\log p(x,s_1|s_2,v_1) - \gamma d\big(x,t(s_2,v_1)\big) - y_{x,s_1,s_2,v_1,t}\Big] \leq 0, \quad \forall x,s_1,v_1,t, \\ & \log\Big(\sum_{x,s_1} \exp\big\{y_{x,s_1,s_2,v_1,,t}\big\}\Big) \leq 0, \quad \forall s_2,v_1,t, \\ & \gamma \geq 0, \end{array} \tag{45}$$

where the variables of the maximization are $\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{X}||\mathcal{S}_1||\mathcal{V}_1|}, \gamma \in \mathbb{R}$ and $\mathbf{y} \in \mathbb{R}^{|\mathcal{X}||\mathcal{S}_1||\mathcal{S}_2||\mathcal{V}_1||\mathcal{T}|}$. It can be verified that this geometric program is a generalization of the geometric program we developed in Theorem 3 and that it corresponds to the problem of minimizing $I(T;X,S_1|V_1) - I(T;S_2|V_1)$ over $q(t|x,s_1,v_1)$ s.t. $\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} d(X_i,\hat{X}_i)\Big] \leq D$ (for a fixed probability $w(v_1|s_1)$). Therefore, all we are left to do now is to declare

$$R(D) = \min_{w(v_1|s_1)\in\mathcal{W}^*} R_w(D). \tag{46}$$

This concludes the summary of the algorithm for solving this example.

The numeric result of the calculation of this rate-distortion function is brought in Figure 11.

VI. Semi-Iterative Algorithm

In this section we provide algorithms that numerically calculate the lower bound on the capacity of Case 2 of the channel coding problems. The calculation of the Gelfand-Pinsker and the Wyner-Ziv problems has been addressed in many papers in the past, including [5], [20], [21] and [22]. All these algorithms are based on Arimoto's [19] and Blahut's [18] algorithms and on the fact that the Wyner-Ziv and the Gelfand-Pinsker problems can be presented as convex optimization problems. On the contrary, our problems are not convex in all of their optimization variables

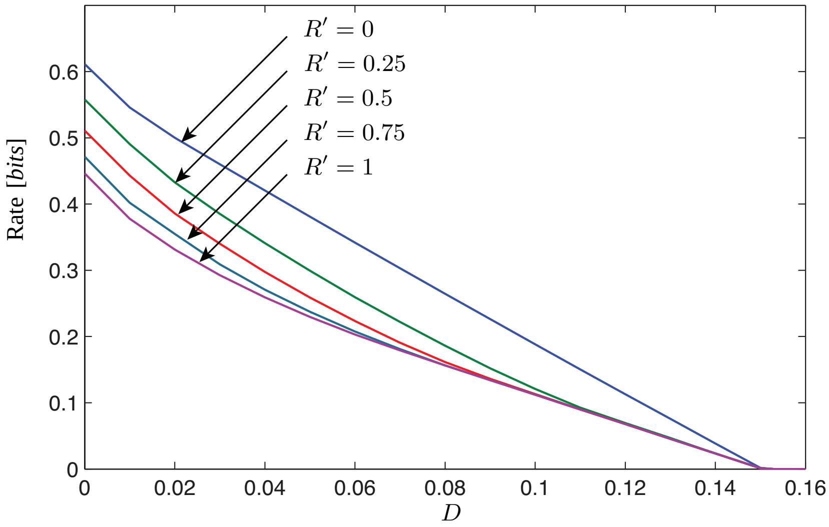

Fig. 11: Example 4. *Geometric programming and source coding Case 1.* The source X is depicted in Figure 10 and the distortion is the Hamming distortion.

and, therefore, cannot be presented as convex optimization problems. In order to solve our problems we devised a different approach which combines a grid-search and a Blauhut-Arimoto-like algorithm. In this section, we provide the mathematical justification for those two algorithms. Other algorithms to numerically compute the channel capacity or the rate-distortion of the rest of the cases presented in this paper can be derived using the principles that we describe in this section.

A. *An algorithm for computing the lower bound on the capacity of Case 2*

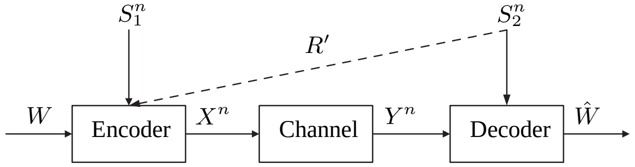

Fig. 12: Channel coding: Case 2. $C_2^{lb} = \max I(U;Y,S_2|V_2) - I(U;S_1|V_2)$, where the maximization is over all PMFs $w(v_2|s_2)p(u|s_1,v_2)p(x|s_1,v_2,u)$ such that $R' \geq I(V_2;S_2|S_1)$.

Consider the channel in Figure 12 described by $p(y|x,s_1,s_2)$ and consider the joint PMF $p(s_1,s_2)$. The capacity of this channel is lower bounded by $\max I(U;Y,S_2|V_2) - I(U;S_1|V_2)$, where the maximization is over all PMFs $p(s_1,s_2)w(v_2|s_2)p(u|s_1,v_2)p(x|s_1,v_2,u)p\big(y|x,s_1,s_2\big)$ such that $R' \geq I(V_2;S_2|S_1)$. Notice that the lower bound expression is not concave in $w(v_2|s_2)$, which is the main difficulty with the computation of it. We first present an outline of the semi-iterative algorithm we developed, then we present the mathematical background and justification for the algorithm and, finally, we present the detailed algorithm.

For any fixed PMF $w(v_2|s_2)$ denote

$$R_w \triangleq I(V_2;S_2|S_1), \tag{47}$$

$$C_{2,w}^{lb} \triangleq \max_{p(u|s_1,v_2)p(x|u,s_1,v_2)} I(U;Y,S_2|V_2) - I(U;S_1|V_2). \tag{48}$$

Then, the lower bound on the capacity , $C_2^{lb}(R')$, can be expressed as

$$C_2^{lb}(R') = \max_{\substack{w(v_2|s_2) \\ \text{s.t. } R' \geq R_w}} \max_{p(u|s_1,v_2)p(x|u,s_1,v_2)} [I(U;Y,S_2|V_2) - I(U;S_1|V_2)] \triangleq \max_{\substack{w(v_2|s_2) \\ \text{s.t. } R' \geq R_w}} C_{2,w}^{lb}. \tag{49}$$

The outline of the algorithm is as follows: for any given rate $R' \leq H(S_2|S_1)$, $\epsilon > 0$ and $\delta > 0$,

1) Establish a fine and uniformly spaced grid of legal PMFs, $w(v_2|s_2)$, and denote the set of all of those PMFs as $\mathcal{W}$.
2) Establish the set $\mathcal{W}^* := \Big\{ w(v_2|s_2) \mid w(v_2|s_2) \in \mathcal{W} \text{ and } R' - \epsilon \leq R_w \leq R' \Big\}$. This set is the set of all PMFs $w(v_2|s_2)$ such that R_w is ϵ-close to R' from below. If $\mathcal{W}^*$ is empty, go back to step 1 and make the grid finer. Otherwise, continue.
3) For every $w(v_2|s_2) \in \mathcal{W}^*$, perform a Blahut-Arimoto-like optimization to find $C_{2,w}^{lb}$ with accuracy of δ.
4) Declare $C_2^{lb}(R') = \max_{w(v_2|s_2) \in \mathcal{W}^*} C_2^{lb(\epsilon,\delta,\mathcal{W})}(R')$.

Remarks: (a) We considered only those R's such that $R' \leq H(S_2|S_1)$ since $H(S_2|S_1)$ is the maximal value that $I(V_2;S_2|S_1)$ takes. The interpretation of this is that if the encoder is informed with S_1, we cannot increase its side information about S_2 in more than $H(S_2|S_1)$. Therefore, for any $H(S_2|S_1) \leq R'$, we can limit R' to be equal to $H(S_2|S_1)$ in order to compute the capacity. (b) Since $C_{2,w}^{lb}(R')$ is continuous in $w(v_2|s_2)$ and bounded (for example, by $I(X;Y|S_1,S_2)$ from above and by $I(X;Y)$ from below), $C_2^{(\epsilon,\delta,\mathcal{W})}(R')$ can be arbitrarily close to $C_2^{lb}(R')$ for $\epsilon \to 0,\ \delta \to 0$ and $|\mathcal{W}| \to \infty$.

Mathematical background and justification

Here we focus on finding the lower bound on the capacity of the channel for a fixed distribution $w(v_2|s_2)$, i.e., finding $C_{2,w}^{lb}$. Note that the mutual information expression $I(U;Y,S_2|V_2) - I(U;S_1|V_2)$ is concave in $p(u|s_1,v_2)$ and convex in $p(x|u,s_1,v_2)$. Therefore, a standard convex maximization technique is not applicable for this problem. However, according to Dupuis, Yu and Willems [21], we can write the expression for the lower bound as $C_{2,w}^{lb} = \max_{q(t|s_1,v_2)} I(T;Y,S_2|V_2) - I(T;S_1|V_2)$, where $q(t|s_1,v_2)$ is a probability distribution over the set of all possible strategies $t : \mathcal{S}_1 \times \mathcal{V}_2 \to \mathcal{X}$, the input symbol X is selected using $x = t(s_1,v_2)$ and $p(y|x,s_1,s_2) = p(y|x,s_1,s_2,v_2) = p\big(y|t(s_1,v_2),s_1,s_2,v_2\big)$. Now, since $I(T;Y,S_2|V_2) - I(T;S_1|V_2)$ is concave in $q(t|s_1,v_2)$, we can use convex optimization methods to derive $C_{2,w}^{lb}$.

Denote the PMF

$$p(s_1,s_2,v_2,t,y) \triangleq p(s_1,s_2)w(v_2|s_2)q(t|s_1,v_2)p(y|t,s_1,s_2,v_2), \tag{50}$$

and denote also

$$J_w(q,Q) \triangleq \sum_{s_1,s_2,v_2,t,y} p(s_1,s_2,v_2,t,y) \log \frac{Q(t|y,s_2,v_2)}{q(t|s_1,v_2)}, \tag{51}$$

$$Q^*(t|y,s_2,v_2) \triangleq \frac{\sum_{s_1} p(s_1,s_2,v_2,t,y)}{\sum_{s_1,t'} p(s_1,s_2,v_2,t',y)}. \tag{52}$$

Notice that $Q^*(t|y,s_2,v_2)$ is a marginal distribution of $p(s_1,s_2,v_2,t,y)$ and that $J_w(q,Q^*) = I(T;Y,S_2|V_2) -$

$I(T;S_1|V_2)$ for the joint PMF $p(s_1,s_2,v_2,t,y)$.

The following lemma is the key for the iterative algorithm.

Lemma 3.

$$C^{lb}_{2,w}=\sup_{q'(t|s_1,v_2)}\max_{Q'(t|y,s_2,v_2)}J_w(q',Q'). \tag{53}$$

The proof for this is brought by Yeung in [29]. In addition, Yeung shows that the two-step alternating optimization procedure converges monotonically to the global optimum if the optimization function is concave. Hence, if we show that $J_w(q,Q)$ is concave, we can maximize it using an alternating maximization algorithm over q and Q.

Lemma 4. The function $J_w(q,Q)$ is concave in q and Q simultaneously.

We can now proceed to calculate the steps in the iterative algorithm.

Lemma 5. For a fixed $q,\ J_w(q,Q)$ is maximized for $Q=Q^*$.

Proof: The above follows from the fact that Q^* is a marginal distribution of $p(s_1,s_2)w(v_2|s_2)q(t|s_1,v_2)$ $p(y|t,s_1,s_2,v_2)$ and the property of the K-L divergence $D(Q^*\|Q')\geq 0$. ■

Lemma 6. For a fixed $Q,\ J_w(q,Q)$ is maximized for $q=q^*$, where q^* is defined by

$$q^*(t|s_1,v_2)=\frac{\prod_{s_2,y}Q(t|y,s_2,v_2)^{p(s_2|s_1,v_2)p(y|t,s_1,s_2,v_2)}}{\sum_{t'}\prod_{s_2,y}Q(t|y,s_2,v_2)^{p(s_2|s_1,v_2)p(y|t',s_1,s_2,v_2)}}, \tag{54}$$

and

$$p(s_2|s_1,v_2)=\frac{p(s_1,s_2)w(v_2|s_2)}{\sum_{s'_2}p(s_1,s'_2)w(v_2|s'_2)}. \tag{55}$$

Define $U_w(q)$ in the following way

$$U_w(q)=\sum_{s_1,v_2}p(s_1,v_2)\max_t\sum_{s_2,y}p(s_2|s_1,v_2)p(y|t,s_1,s_2,v_2)\log\frac{Q^*(t|y,s_2,v_2)}{q(t|s_1,v_2)}, \tag{56}$$

where Q^* is given in (52), $p(s_1,v_2)$ and $p(s_2|s_1,v_2)$ are marginal distributions of the joint PMF $p(s_1,s_2,v_2,t,y)=p(s_1,s_2)w(v_2|s_2)q(t|s_1,v_2)p(y|t,s_1,s_2,v_2)$. The following lemma will help us to define a termination condition for the algorithm.

Lemma 7. For every $q(t|s_1,v_2)$ the function $U_w(q)$ is an upper bound on $C^{lb}_{w,2}$ and converges to $C^{lb}_{2,w}$ for a large enough number of iterations.

B. Semi-iterative algorithm

The the algorithm for finding $C^{lb}_2(R')$ is brought in Algorithm 1. Notice that the result of this algorithm, $C_2^{(\epsilon,\delta,\mathcal{W})}(R')$, can be arbitrarily close to $C^{lb}_2(R')$ for $\epsilon\to 0,\ \delta\to 0$ and $|\mathcal{W}|\to\infty$.

Algorithm 1 Numerically calculating $C_2^{lb}(R')$

```
1: Chose $\epsilon > 0$, $\delta > 0$
2: Set $R' \leftarrow \min\{R', H(S_2|S_1)\}$    ▷ the amount of information needed for the encoder to know $S_2$ given $S_1$
3: Set $C \leftarrow -\infty$
4: Establish a fine and uniformly spaced grid of legal PMFs $w(v_2|s_2)$ and name it $\mathcal{W}$
5: for all $w$ in $\mathcal{W}$ do
6:     Compute $R_w$ using
           $$R_w = I(V_2;S_2) - I(V_2;S_1)$$
7:     if $R' - \epsilon \leq R_w \leq R'$ then
8:         Set $Q(t|y,s_2,v_2)$ to be a uniform distribution over $\{1,2,\dots,|\mathcal{T}|\}$, where $\mathcal{T}$ is the alphabet of $t$.
           i.e., $Q(t|y,s_2,v_2) = \frac{1}{|\mathcal{T}|}, \quad \forall t,y,s_2,v_2$
9:         repeat
10:            Set $q(t|s_1,v_2) \leftarrow q^*(t|s_1,v_2)$ using
               $$q^*(t|s_1,v_2) = \frac{\prod_{s_2,y} Q(t|y,s_2,v_2)^{p(s_2|s_1,v_2)p(y|t,s_1,s_2,v_2)}}{\sum_{t'}\prod_{s_2,y} Q(t'|y,s_2,v_2)^{p(s_2|s_1,v_2)p(y|t',s_1,s_2,v_2)}}$$
11:            Set $(Q(t|y,s_2,v_2) \leftarrow Q^*(t|y,s_2,v_2)$ using
               $$Q^*(t|y,s_2,v_2) = \frac{\sum_{s_1} p(s_1,s_2,v_2,t,y)}{\sum_{s_1,t'} p(s_1,s_2,v_2,t',y)}$$
12:            Compute $J_w(q,Q)$ using
               $$J_w(q,Q) = \sum_{s_1,s_2,v_2,t,y} p(s_1,s_2,v_2,t,y)\log\frac{Q(t|y,s_2,v_2)}{q(t|s_1,v_2)}$$
13:            Compute $U_w(q)$ using
               $$U_w(q) = \sum_{s_1,v_2} p(s_1,v_2)\max_t \sum_{s_2,y} p(s_2|s_1,v_2)p(y|t,s_1,s_2,v_2)\log\frac{Q^*(t|y,s_2,v_2)}{q(t|s_1,v_2)}$$
14:        until $U_w(q) - J(q,Q) < \delta$
15:        if $C \leq J_w(q,Q)$ then
16:            Set $C \leftarrow J_w(q,Q)$
17:        end if
18:    end if
19: end for
20: if $C < 0$ then    ▷ there is no PMF $w(v_2|s_2) \in \mathcal{W}$ such that $R_w$ is $\epsilon$-close to $R'$ from below
21:    go to line 4 and make the grid finer
22: end if
23: Declare $C_2^{lb(\epsilon,\delta,\mathcal{W})}(R') = C$
```

VII. Open Problems

In this section we discuss the generalization of the channel capacity and the rate-distortion problems that we presented in Section III. We now consider the cases where the encoder and the decoder are informed with both a rate-limited description of the ESI and a rate-limited description of the DSI simultaneously, as illustrated in Figure 13. Although proofs for the converses are not provided in this paper and are considered as open problems, we do provide achievability schemes for both problems.

A. A lower bound on the capacity of a channel with two-sided increased partial side information

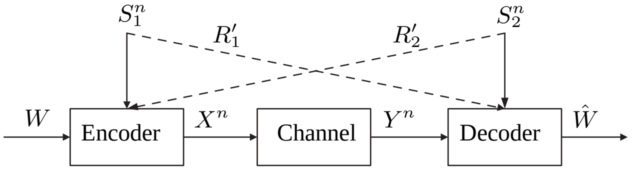

Fig. 13: A lower bound on the capacity of a channel with two-sided increased partial side information: $C_{12} \geq \max I(U;Y,S_2|V_1,V_2) - I(U;S_1|V_1,V_2)$, where the maximization is over all PMFs $p(v_1|s_1)p(v_2|s_2)p(u|s_1,v_1,v_2)p(x|u,s_1,v_1,v_2)$ such that $R_1' \geq I(V_1;S_1) - I(V_1;Y,S_2,V_2)$ and $R_2' \geq I(V_2;S_2) - I(V_2;S_1,V_1)$.

Consider the channel illustrated in Figure 13, where $(S_{1,i},S_{2,i})$ i.i.d. $\sim p(s_1,s_2)$. The encoder is informed with the ESI (S_1^n) and rate-limited DSI and the decoder is informed with the DSI (S_2^n) and rate-limited ESI. An $(n,2^{nR},2^{nR_1'},2^{nR_2'})$ code for the discussed channel consists of three encoding maps:

$$\begin{aligned}
f_{v1}: &\quad \mathcal{S}_1^n \mapsto \{1,2,\dots,2^{nR_1'}\},\\
f_{v2}: &\quad \mathcal{S}_2^n \mapsto \{1,2,\dots,2^{nR_2'}\},\\
f: &\quad \{1,2,\dots,2^{nR}\} \times \mathcal{S}_1^n \times \{1,2,\dots,2^{nR_2'}\} \mapsto \mathcal{X}^n,
\end{aligned}$$

and a decoding map:

$$g: \mathcal{Y}^n \times \mathcal{S}_2^n \times \{1,2,\dots,2^{nR_1'}\} \mapsto \{1,2,\dots,2^{nR}\}.$$

Fact 1: The channel capacity, C_{12}^*, of this channel coding setup is bounded from below as follows:

$$C_{12}^* \geq \max_{\substack{p(v_1|s_1)p(v_2|s_2)p(u|s_1,v_1,v_2)p(x|u,s_1,v_1,v_2)\\ \text{s.t.}\quad R_1' \geq I(V_1;S_1)-I(V_1;Y,S_2,V_2)\\ R_2' \geq I(V_2;S_2)-I(V_2;S_1)}} I(U;Y,S_2|V_1,V_2) - I(U;S_1|V_1,V_2), \tag{57}$$

for some joint distribution $p(s_1,s_2,v_1,v_2,u,x,y)$ and U,V_1 and V_2 are some auxiliary random variables.

The proof for the achievability follows closely the proofs given in Appendix B and, therefore, we only provide the outline of the achievability. The main steps of the achievability scheme are outlined in the following.

Sketch of proof of Achievability for Fact 1: (a) *The ESI encoder wants to describe S_1^n to the decoder with rate of R_1'.* We generate $2^{n(I(V_1;S_1)+\epsilon)}$ sequences V_1^n i.i.d. $\sim p(v_1)$ and randomly distribute them into $2^{n\left(I(V_1;S_1)-I(V_1;Y,S_2,V_2)+2\epsilon\right)}$ bins; each bin contains $2^{n(I(V_1;Y,S_2,V_2)-\epsilon)}$ codewords. The ESI encoder is given the sequence s_1^n and first looks for a sequence v_1^n that is jointly typical with s_1^n. If there is such a codeword, the ESI encoder sends the index of the bin that contains v_1^n to the decoder. The decoder, given y^n, s_2^n, v_2^n, looks for a unique codeword in the received bin that is jointly typical with y^n, s_2^n, v_2^n. Since there are more than $2^{nI(V_1;S_1)}$ sequences V_1^n, the ESI encoder is assured with high probability to find a sequence v_1^n such that $(v_1^n, s_1^n) \in \mathcal{T}_\epsilon^{(n)}(V_1, S_1)$. Since, in addition, there are less than $2^{nI(V_1;Y,S_2,V_2)}$ codewords in the bin, the decoder is assured to find a unique sequence v_1^n in the bin such that $(v_1^n, y^n, s_2^n, v_2^n) \in \mathcal{T}_\epsilon^{(n)}(V_1, Y, S_2, V_2)$ with high probability. Therefore, the constraint on the shared ESI is maintained if $R_1' > I(V_1; S_1) - I(V_1; Y, S_2, V_2)$.

(b) *The DSI encoder wants to describe S_2^n to the channel's encoder with a rate of R_2'*. We generate $2^{n(I(V_2;S_2)+\epsilon)}$ sequences $V_2^n \sim$ i.i.d. $p(v_2)$ and randomly distribute them into $2^{n\left(I(V_2;S_2)-I(V_2;S_1,V_1)+2\epsilon\right)}$ bins; each bin contains $2^{n(I(V_2;S_1,V_1)-\epsilon)}$ codewords. The DSI encoder, given s_2^n, first looks for a sequence v_2^n that is jointly typical with s_2^n. If there is such a codeword, the DSI encoder sends the index of the bin where v_2^n is located to the channel's encoder. The channel's encoder, given s_1^n, v_1^n, looks for a unique sequence v_2^n in the received bin that is jointly typical with s_1^n, v_1^n. Since there are more than $2^{nI(V_2;S_2)}$ sequences V_2^n, the DSI encoder is assured with high probability to find such a sequence v_2^n such that $(v_2^n, s_2^n) \in \mathcal{T}_\epsilon^{(n)}(V_2, S_2)$. In its turn, the channel's encoder is also assured with high probability to find the unique sequence v_2^n in its received bin such that $(v_2^n, s_1^n, v_1^n) \in \mathcal{T}_\epsilon^{(n)}(V_2, S_1, V_1)$, since there are less than $2^{nI(V_2;S_1,V_1)}$ codewords V_2^n in the bin. Therefore, the constraint of the shared DSI is maintained if $R_2' > I(V_2; S_2) - I(V_2; S_1, V_1)$.

(c) *The encoder wants to send the message W to the decoder*. For each v_1^n, v_2^n we generate $2^{n(I(U;Y,S_2|V_1,V_2)-\epsilon)}$ sequences U^n using the PMF $p(u^n|v_1^n, v_2^n) = \prod_{i=1}^n p(u_i|v_{1,i}, v_{2,i})$ and randomly distribute them into $2^{n\left(I(U;Y,S_2,|V_1,V_2)-I(U;S_1|V_1,V_2)-2\epsilon\right)}$ bins; each bin contains $2^{n(I(U;S_1|V_1,V_2)+\epsilon)}$ codewords. The encoder, given s_1^n, v_1^n, v_2^n and the message W, looks in the bin number W for a sequence u^n that is jointly typical with s_1^n, v_1^n, v_2^n and sends $x_i = f(u_i, s_{1,i}, v_{1,i}, v_{2,i})$ over the channel at time i. The decoder receives y^n, s_2^n, v_1^n, v_2^n and first looks for a unique sequence u^n that is jointly typical with y^n, s_2^n, v_1^n, v_2^n. Upon finding the desired sequence u^n, the decoder declares $\hat{W}$ to be the index of the bin that contains u^n. Having less than $2^{nI(U;Y,S_2|V_1,V_2)}$ sequences U^n assures with high probability that decoder will identify a unique sequence u^n such that $(u^n, y^n, s_2^n, v_1^n, v_2^n) \in \mathcal{T}_\epsilon^{(n)}(U, Y, S_2, v_1, v_2)$. This is also valid because the Markov relation $(U, V_1, V_2) - (X, S_1, S_2) - Y$ implies that $(u^n, v_1^n, v_2^n, x^n, s_1^n, s_2^n, y^n) \in \mathcal{T}_\epsilon^{(n)}(U, V_1, V_2, X, S_1, S_2, Y)$. In addition, since in each of the encoder's bins there are more than $2^{nI(U;S_1|V_1,V_2)}$ codewords U^n, the encoder is assured with high probability to find a sequence u^n in the bin indexed W such that $(u^n, s_1^n, v_1^n, v_2^n) \in \mathcal{T}_\epsilon^{(n)}(U, S_1, v_1, v_2)$. We can conclude that if $R < I(U; Y, S_2|V_1, V_2) - I(U; S_1|V_1, V_2)$ is maintained, then a reliable communication over the channel is achievable; namely, it is possible to find a sequence of codes such the $\Pr\{\hat{W} \neq W\}$ goes to zero as the block length goes to infinity. This concludes the sketch of the achievability.

B. An upper bound on the rate-distortion with two-sided increased partial side information

Consider the rate-distortion problem illustrated in Figure 14, where the source X and the side information S_1, S_2 are distributed $(X_i, S_{1,i}, S_{2,i}) \sim$ i.i.d. $p(x, s_1, s_2)$. The encoder is informed with the ESI (S_1^n) and rate-limited

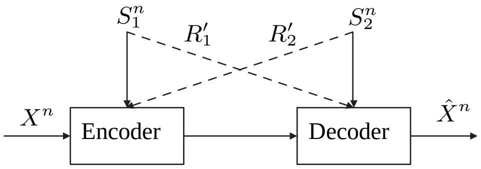

Fig. 14: An upper bound on the rate-distortion with two-sided increased partial side information: $R_{12}(D) \leq \min I(U;X,S_1|V_1,V_2) - I(U;S_2|V_1,V_2)$, where the minimization is over all PMFs $p(v_1|s_1)p(v_2|s_2)p(u|x,s_1,v_1,v_2)p(\hat{x}|u,s_2,v_1,v_2)$ such that $R_1' \geq I(V_1;S_1) - I(V_1;S_2)$, $R_2' \geq I(V_2;S_2) - I(V_2;X,S_1,V_1)$ and $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} d(X,\hat{X})\right] \leq D$.

DSI and the decoder is informed with the DSI (S_2^n) and rate-limited ESI. An $(n, 2^{nR}, 2^{nR_1'}, 2^{nR_2'}, D)$ code for the discussed rate-distortion problem consists of three encoding maps:

$$\begin{aligned} f_{v1}: &\quad \mathcal{S}_1^n \mapsto \{1,2,\dots,2^{nR_1'}\}, \\ f_{v2}: &\quad \mathcal{S}_2^n \mapsto \{1,2,\dots,2^{nR_2'}\}, \\ f: &\quad \mathcal{X}^n \times \mathcal{S}_1^n \times \{1,2,\dots,2^{nR_2'}\} \mapsto \{1,2,\dots,2^{nR}\}, \end{aligned}$$

and a decoding map:

$$g: \{1,2,\dots,2^{nR}\} \times \mathcal{S}_2^n \times \{1,2,\dots,2^{nR_1'}\} \mapsto \hat{\mathcal{X}}^n.$$

Fact 2: For a given distortion, D, and a given distortion measure, $d(X,\hat{X}): \quad \mathcal{X} \times \hat{\mathcal{X}} \mapsto \mathbb{R}^+$, the rate-distortion function $R_{12}^*(D)$ of this setup is bounded from above as follows:

$$R_{12}^*(D) \leq \min_{\substack{p(v_1|s_1)p(v_2|s_2)p(u|x,s_1,v_1,v_2)p(\hat{x}|u,s_2,v_1,v_2) \\ \text{s.t. } R_1' \geq I(V_1;S_1) - I(V_1;S_2,V_2) \\ R_2' \geq I(V_2;S_2) - I(V_2;X,S_1,V_1)}} I(U;X,S_1|V_1,V_2) - I(U;S_2|V_1,V_2), \tag{58}$$

for some joint distribution $p(x,s_1,s_2,v_1,v_2,u,\hat{x})$ where $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} d(X_i,\hat{X}_i)\right] \leq D$ and U, V_1 and V_2 are some auxiliary random variables.

The achievability proof is outlined in the following. The steps of the proof resemble the steps of the achievability proof for Fact 1.

Sketch of proof of Achievability for Fact 2: (a) *The ESI encoder wants to describe S_1^n to the decoder with a rate of R_1'.* We generate $2^{n(I(V_1;S_1)+\epsilon)}$ sequences V_1^n i.i.d. $\sim p(v_1)$ and randomly distribute them into $2^{n\left(I(V_1;S_1)-I(V_1;S_2,V_2)+2\epsilon\right)}$ bins; each bin contains $2^{n(I(V_1;S_2,V_2)-\epsilon)}$ codewords. The ESI encoder is given the sequence s_1^n and first looks for a sequence v_1^n that is jointly typical with s_1^n. If there is such a codeword, the ESI encoder sends the index of the bin that contains v_1^n to the decoder. The decoder, given s_2^n, v_2^n, looks for a unique codeword in the received bin that is jointly typical with s_2^n, v_2^n. Since there are more than $2^{nI(V_1;S_1)}$ sequences V_1^n, the ESI encoder is assured with high probability to find a sequence v_1^n such that $(v_1^n, s_1^n) \in \mathcal{T}_\epsilon^{(n)}(V_1,S_1)$. Since, in addition, there are less than $2^{nI(V_1;S_2,V_2)}$ codewords in the bin, the decoder is assured with high probability to find a unique sequence v_1^n in the bin such that $(v_1^n s_2^n, v_2^n) \in \mathcal{T}_\epsilon^{(n)}(V_1,S_2,V_2)$. Therefore, the constraint on the rate of the shared ESI is maintained if $R_1' > I(V_1;S_1) - I(V_1;S_2,V_2)$.

(b) *The DSI encoder wants to describe S_2^n to the source encoder with a rate of R_2'.* We generate $2^{n(I(V_2;S_2)+\epsilon)}$ sequences $V_2^n \sim$ i.i.d. $p(v_2)$ and randomly distribute them into $2^{n\left(I(V_2;S_2)-I(V_2;X,S_1,V_1)+2\epsilon\right)}$ bins; each bin contains $2^{n(I(V_2;X,S_1,V_1)-\epsilon)}$ codewords. The DSI encoder, given s_2^n, first looks for a sequence v_2^n that is jointly typical with s_2^n. If there is such a codeword, the DSI encoder sends the index of the bin where v_2^n is located to the source encoder. The source encoder, given x^n, s_1^n, v_1^n, looks for a unique sequence v_2^n in the received bin that is jointly typical with x^n, s_1^n, v_1^n. Since there are more than $2^{nI(V_2;S_2)}$ sequences V_2^n, the DSI encoder is assured with high probability to find a sequence v_2^n such that $(v_2^n, s_2^n) \in \mathcal{T}_\epsilon^{(n)}(V_2, S_2)$. At the same time, the source encoder is assured with high probability to find the unique sequence v_2^n in its received bin such that $(v_2^n, x^n, s_1^n, v_1^n) \in \mathcal{T}_\epsilon^{(n)}(V_2, X, S_1, V_1)$, since there are less than $2^{nI(V_2;X,S_1,V_1)}$ codewords V_2^n in the bin. Therefore, the constraint on the rate of the shared DSI is maintained if $R_2' > I(V_2;S_2) - I(V_2;X,S_1,V_1)$.

(c) *The source encoder wants to describe the source X to the decoder with distortion smaller than or equal to D; that is* $\mathbb{E}\left[d(X,\hat{X})\right] \leq D$. For each v_1^n, v_2^n we generate $2^{n(I(U;X,S_1|V_1,V_2)+\epsilon)}$ sequences U^n using the PMF $p(u^n|v_1^n, v_2^n) = \prod_{i=1}^n p(u_i|v_{1,i}, v_{2,i})$ and randomly distribute them into $2^{n\left(I(U;X,S_1,|V_1,V_2)-I(U;S_2|V_1,V_2)+2\epsilon\right)}$ bins; each bin contains $2^{n(I(U;S_2|V_1,V_2)-\epsilon)}$ codewords. The source encoder, given x^n, s_1^n, v_1^n, v_2^n, looks for a sequence u^n that is jointly typical with x^n, s_1^n, v_1^n, v_2^n and sends the index of the bin that contains u^n to the decoder. The decoder, given s_2^n, v_1^n, v_2^n, looks for a unique sequence u^n in the received bin that is jointly typical with s_2^n, v_1^n, v_2^n. Upon finding the desired sequence u^n, the decoder declares $\hat{x}_i = g(u_i, s_{2,i}, v_{1,i}, v_{2,i})$ for $i \in \{1, 2, \ldots, n\}$ to be the reconstruction of the source x^n. Having more than $2^{nI(U;X,S_1|V_1,V_2)}$ sequences U^n assures the encoder with high probability to find a sequence u^n such that $(u^n, x^n, s_1^n, v_1^n, v_2^n) \in \mathcal{T}_\epsilon^{(n)}(U, X, S_1, v_1, v_2)$. Since, in addition, each one of the bins contains there are less than $2^{nI(U;S_2|V_1,V_2)}$ codewords U^n, the decoder is assured with high probability to find a unique sequence u^n in the bin such that $(u^n, s_2^n, v_1^n, v_2^n) \in \mathcal{T}_\epsilon^{(n)}(U, S_2, v_1, v_2)$. Therefore, and since the Markov chain $(X, S_1) - (U, S_2, V_1, V_2) - \hat{X}$ is satisfied, we can conclude that a rate of $R > I(U;X,S_1|V_1,V_2) - I(U;S_2|V_1,V_2)$ allows the decoder to produce $\hat{x}^n$ that satisfies the distortion constraint with high probability; i.e., that $d(x^n, \hat{x}^n) \leq D$ with high probability. This concludes the sketch of the proof of the achievability.

Appendix A

Duality of the Converse of the Gelfand-Pinsker Theorem and the Wyner-Ziv Theorem

In this appendix we provide proofs of the converse of the Gelfand-Pinsker capacity and the converse of the Wyner-Ziv rate in a dual way.

	Channel capacity	Rate-distortion
1	$nR \quad = H(W)$	$nR \quad = H(T)$
2	$\overset{(a)}{\leq} I(W;Y^n) - I(W;S^n) + n\epsilon_n$	$\overset{(a)}{\geq} I(T;X^n) - I(T;S^n)$
3	$= \sum_{i=1}^n \Big[I(W;Y_i\|Y^{i-1})$ $-I(W;S_i\|S_{i+1}^n)\Big] + n\epsilon_n$	$= \sum_{i=1}^n \Big[I(T;X_i\|X^{i-1})$ $-I(T;S_i\|S_{i+1}^n)\Big]$
4	$= \sum_{i=1}^n \Big[I(W,S_{i+1}^n;Y_i\|Y^{i-1})$ $-I(W,Y^{i-1};S_i\|S_{i+1}^n)\Big] + \Delta - \Delta^* + n\epsilon_n$	$= \sum_{i=1}^n \Big[I(T,S_{i+1}^n;X_i\|X^{i-1})$ $-I(W,X^{i-1};S_i\|S_{i+1}^n)\Big] + \Delta - \Delta^*$
5	$\overset{(b)}{\leq} \sum_{i=1}^n \Big[I(W,,Y^{i-1},S_{i+1}^n;Y_i)$ $-I(W,Y^{i-1},S_{i+1}^n;S_i)\Big] + n\epsilon_n$	$\overset{(b)}{\geq} \sum_{i=1}^n \Big[I(T,,X^{i-1},S_{i+1}^n;X_i)$ $-I(T,X^{i-1},S_{i+1}^n;S_i)\Big]$
6	$= \sum_{i=1}^n \Big[I(U_i;Y_i) - I(U_i;S_i)\Big] + n\epsilon_n,$	$= \sum_{i=1}^n \Big[I(U_i;X_i) - I(U_i;S_i)\Big],$

(59)

where

Δ	$= \sum_{i=1}^n I(Y^{i-1};S_i\|W,S_{i+1}^n),$	Δ	$= \sum_{i=1}^n I(X^{i-1};S_i\|T,S_{i+1}^n),$
Δ^*	$= \sum_{i=1}^n I(S_{i+1}^n;Y_i\|W,Y^{i-1}),$	Δ^*	$= \sum_{i=1}^n I(S_{i+1}^n;X_i\|T,X^{i-1}),$
(a)	follows from Fano's inequality and from that fact that W is independent of S^n,	(a)	follows from Fano's inequality and from the fact that T is independent of S^n,
(b)	follows from the fact that S_i is independent of S_{i+1}^n.	(b)	follows from the fact that S_i is independent of S_{i+1}^n and that X_i is independent of X^{i-1}.

(60)

By substituting the output Y and the input X in the channel capacity theorem with the input X and the output $\hat{X}$ in the rate-distortion theorem, respectively, we can observe duality in the converse proofs of the two theorems.

APPENDIX B
PROOF OF THEOREM 1

In this section we provide the proofs for Theorem 1, Cases 2 and 2_C. The results for Case 1, where the encoder is informed with ESI and the decoder is informed with increased DSI, can be derived directly from [2, Theorem VII]. In [2], Steinberg considered the case where the encoder is fully informed with the ESI and the decoder is informed with a rate-limited description of the ESI. Therefore, by considering the DSI, S_2^n, to be a part of the channel's output, we can apply Steinberg's result on the channel depicted in Case 1. For this reason, the proof for this case is omitted.

A. Proof of Theorem 1, Case 2

The proof of the lower bound, C_2^{lb}, is performed in the following way: for the description of the DSI, S_2, at a rate R' we use a Wyner-Ziv coding scheme where the source is S_2 and the side information is S_1. Then, for the

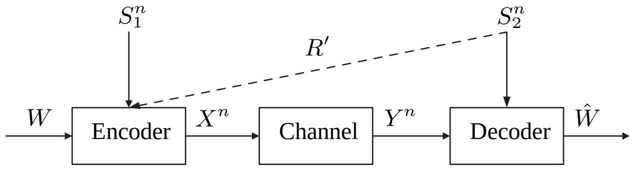

Fig. 15: Channel capacity: Case 2. Lower bound: $C_2^{lb} = \max I(U;Y,S_2|V_2) - I(U;S_1|V_2)$, where the maximization is over all joint PMFs $p(s_1,s_2,v_2,u,x,y)$ that maintain the Markov relations $U - (S_1,V_2) - S_2$ and $V_2 - S_2 - S_1$ and the constraint $R' \geq I(V_2;S_2|S_1)$. Upper bounds: C_2^{ub1} is the result of the same expressions as for the lower bound, except that the maximization is taken over all PMFs that maintain the Markov chain $U - (S_1,V_2) - S_2$, and C_2^{ub2} is the result of the same expressions as for the lower bound, except that this time the maximization is taken over all PMFs that maintain $V_2 - S_2 - S_1$.

channel coding, we use a Gelfand-Pinsker coding scheme where the state information at the encoder is S_1, S_2 is a part of the channel's output and the rate-limited description of S_2 is side information at both the encoder and the decoder. Notice that $I(U;Y,S_2|V_2) - I(U;S_1,|V_2) = I(U;Y,S_2,V_2) - I(U;S_1,V_2)$ and that, since the Markov chain $V_2 - S_2 - S_1$ holds, we can also write $R' \geq I(V_2;S_2) - I(V_2;S_1)$. We make use of these expressions in the following proof.

Achievability: (*Channel capacity Case 2 - Lower bound*). Given $(S_{1,i}, S_{2,i}) \sim$ i.i.d. $p(s_1,s_2)$ and the memoryless channel $p(y|x,s_1,s_2)$, fix $p(s_1,s_2,v_2,u,x,y) = p(s_1,s_2)p(v_2|s_2)p(u|s_1,v_2)p(x|u,s_1,v_2)p(y|x,s_1,s_2)$, where $x = f(u,s_1,v_2)$ (i.e., $p(x|u,s_1,v_2)$ can get the values 0 or 1).

Codebook generation and random binning

1) Generate a codebook $\mathcal{C}_v$ of $2^{n(I(V_2;S_2))+2\epsilon}$ sequences V_2^n independently using i.i.d. $\sim p(v_2)$. Label them $v_2^n(k)$, where $k \in \{1,2,\dots,2^{n(I(V_2;S_2)+2\epsilon)}\}$, and randomly assign each sequence $v_2^n(k)$ a bin number $b_v\big(v_2^n(k)\big)$ in the set $\{1,2,\dots,2^{nR'}\}$.
2) Generate a codebook $\mathcal{C}_u$ of $2^{n(I(U;Y,S_2,V_2)-2\epsilon)}$ sequences U^n independently using i.i.d. $\sim p(u)$. Label them $u^n(l)$, $l \in \{1,2,\dots,2^{n(I(U;Y,S_2,V_2)-2\epsilon)}\}$, and randomly assign each sequence a bin number $b_u\big(u^n(l)\big)$ in the set $\{1,2,\dots,2^{nR}\}$.

Reveal the codebooks and the content of the bins to all encoders and decoders.

Encoding

1) *State Encoder*: Given the sequence S_2^n, search the codebook $\mathcal{C}_v$ and identify an index k such that $\big(v_2^n(k), S_2^n\big) \in \mathcal{T}_\epsilon^{(n)}(V_2,S_2)$. If such a k is found, stop searching and send the bin number $j = b_v\big(v_2^n(k)\big)$. If no such k is found, declare an error.
2) *Encoder*: Given the message W, the sequence S_1^n and the index j, search the codebook $\mathcal{C}_v$ and identify an index k such that $\big(v_2^n(k), S_1\big) \in \mathcal{T}_\epsilon^{(n)}(V_2,S_1)$. If no such k is found or there is more than one such index, declare an error. If a unique k, as defined, is found, search the codebook $\mathcal{C}_u$ and identify an index l such that $\big(u^n(l), S_1^n, v_2^n(k)\big) \in \mathcal{T}_\epsilon^{(n)}(U,S_1,V_2)$ and $b_u\big(u^n(l)\big) = W$. If a unique l, as defined, is found, transmit $x_i = f\big(u_i(l), S_{1,i}, v_{2,i}(k)\big)$, $i = 1,2,\dots,n$. Otherwise, if there is no such l or there is more than one, declare an error.

Decoding

Given the sequences Y^n, S_2^n and the index k, search the codebook $\mathcal{C}_u$ and identify an index l such that $\big(u^n(l), Y^n, S_2^n, v_2^n(k)\big) \in \mathcal{T}_\epsilon^{(n)}(U, Y, S_2, V_2)$. If a unique l, as defined, is found, declare the message $\hat{W}$ to be the bin index where $u^n(l)$ is located, i.e., $\hat{W} = b_u\big(u^n(l)\big)$. Otherwise, if no such l is found or there is more than one, declare an error.

Analysis of the probability of error

Without loss of generality, let us assume that the message $W = 1$ was sent and the indexes that correspond with the given $W = 1, S_1^n, S_2^n$ are $(k = 1, l = 1$ and $j = 1)$; i.e., $v_2^n(1)$ corresponds with S_2^n, $b_v\big(v_2^n(1)\big) = 1$, $u^n(1)$ is chosen according to $\big(W = 1, S_1^n, v_2^n(1)\big)$ and $b_u\big(u^n(1)\big) = 1$.

Define the following events:

$$
\begin{aligned}
&E_1 := \Big\{\forall v_2^n(k) \in \mathcal{C}_v,\ \big(v_2^n(k), S_2^n\big) \notin \mathcal{T}_\epsilon^{(n)}(V_2, S_2)\Big\}\\
&E_2 := \Big\{\big(v_2^n(1), S_1^n\big) \notin \mathcal{T}_\epsilon^{(n)}(V_2, S_1)\Big\}\\
&E_3 := \Big\{\exists k' \neq 1 \text{ such that } b_v\big(v_2^n(k')\big) = 1 \text{ and } \big(v_2^n(k'), S_1^n\big) \in \mathcal{T}_\epsilon^{(n)}(V_2, S_1)\Big\}\\
&E_4 := \Big\{\forall u^n(l) \in \mathcal{C}_u \text{ such that } b_u\big(u^n(l)\big) = 1,\ \big(u^n(l), S_1^n, v_2^n(1)\big) \notin \mathcal{T}_\epsilon^{(n)}(U, S_1, V_2)\Big\}\\
&E_5 := \Big\{\big(u^n(1), Y^n, S_2^n, v_2^n(1)\big) \notin \mathcal{T}_\epsilon^{(n)}(U, Y, S_2, V_2)\Big\}\\
&E_6 := \Big\{\exists l' \neq 1 \text{ such that } \big(u^n(l'), Y^n, S_2^n, v_2^n(1)\big) \in \mathcal{T}_\epsilon^{(n)}(U, Y, S_2, V_2)\Big\}
\end{aligned}
$$

The probability of error $P_e^{(n)}$ is upper bounded by $P_e^n \leq P(E_1) + P(E_2|E_1^c) + P(E_3|E_1^c, E_2^c) + P(E_4|E_1^c, E_2^c, E_3^c) + P(E_5|E_1^c, \ldots, E_4^c) + P(E_6|E_1^c, \ldots, E_5^c)$. Using standard arguments, and assuming that $(S_1^n, S_2^n) \in \mathcal{T}_\epsilon^{(n)}(S_1, S_2)$ and that n is large enough, we can state that

1)

$$
\begin{aligned}
P(E_1) =& \Pr\Big\{ \bigcap_{v_2^n(k) \in \mathcal{C}_v} \big(v_2^n(k), S_2^n\big) \notin \mathcal{T}_\epsilon^{(n)}(V_2, S_2)\Big\}\\
=& \prod_{k=1}^{2^{n(I(V_2;S_2)+2\epsilon)}} \Pr\Big\{\big(v_2^n(k), S_2^n\big) \notin \mathcal{T}_\epsilon^{(n)}(V_2, S_2)\Big\}\\
=& \prod_{k=1}^{2^{n(I(V_2;S_2)+2\epsilon)}} \Big(1 - \Pr\Big\{\big(v_2^n(k), S_2^n\big) \in \mathcal{T}_\epsilon^{(n)}(V_2, S_2)\Big\}\Big)\\
\leq& \Big(1 - 2^{-n(I(V_2;S_2)+\epsilon)}\Big)^{2^{n(I(V_2;S_2)+2\epsilon)}}\\
\leq& e^{-2^{-n(I(V_2;S_2)+\epsilon)} 2^{n(I(V_2;S_2)+2\epsilon)}}\\
=& e^{-2^{n\epsilon}}.
\end{aligned}
\tag{61}
$$

The probability that there is no $v_2^n(k)$ in $\mathcal{C}_v$ such that $\big(v_2^n(k), S_2^n\big)$ is strongly jointly typical is exponentially small provided that $|\mathcal{C}_v| \geq 2^{n(I(V_2;S_2)+\epsilon)}$. This follows from the standard rate-distortion argument that

$2^{nI(V_2;S_2)}$ v_2^n's “cover” $\mathcal{S}_2^n$, therefore $P(E_1) \mapsto 0$.

2) By the Markov lemma [30], since (S_1^n, S_2^n) are strongly jointly typical, $\big(S_2^n, v_2^n(1)\big)$ are strongly jointly typical and the Markov chain $S_1 - S_2 - V_2$ holds, then $\big(S_1^n, S_2^n, v_2^n(1)\big)$ are strongly jointly typical with high probability. Therefore, $P(E_2|E_1^c) \to 0$.

3)

$$\begin{align}
P(E_3|E_1^c, E_2^c) &= \Pr\Big\{ \bigcup_{\substack{v_2^n(k'\neq 1)\in\mathcal{C}_v \\ b_v\left(v_2^n(k')\right)=1}} \big(v_2^n(k'), S_1^n\big) \in \mathcal{T}_\epsilon^{(n)}(V_2, S_1) \Big\} \tag{62}\\
&\leq \sum_{\substack{v_2^n(k'\neq 1)\in\mathcal{C}_v \\ b_v\left(v_2^n(k')\right)=1}} \Pr\Big\{ \big(v_2^n(k'), S_1^n\big) \in \mathcal{T}_\epsilon^{(n)}(V_2, S_1) \Big\} \tag{63}\\
&\leq \sum_{\substack{v_2^n(k'\neq 1)\in\mathcal{C}_v \\ b_v\left(v_2^n(k')\right)=1}} 2^{n(I(V_2;S_1)+\epsilon)} \tag{64}\\
&= 2^{n(I(V_2;S_2)+2\epsilon-R')} 2^{-n(I(V_2;S_1)-\epsilon)} \tag{65}\\
&= 2^{n(I(V_2;S_2)-I(V_2;S_1)+3\epsilon-R')}. \tag{66}
\end{align}$$

The probability that there is another index k', $k' \neq 1$, such that $v_2^n(k')$ is in bin number 1 and that is strongly jointly typical with S_1^n is bounded by the number of $v_2^n(k')$'s in the bin times the probability of joint typicality. Therefore, if the number of bins $R' > I(V_2;S_2) - I(V_2;S_1) + 3\epsilon$ then $P(E_3|E_1^c, E_2^c) \to 0$.

4) We use here the same argument we used for $P(E_1)$; by the covering lemma, we can state that the probability that there is no $u^n(l)$ in bin number 1 that is strongly jointly typical with $\big(S_1^n, v_2^n(1)\big)$ tends to zero for large enough n if the average number of $u^n(l)$'s in each bin is greater than $2^{n(I(U;S_1,V_2)+\epsilon)}$; i.e., $|\mathcal{C}_u|/2^{nR} > 2^{n(I(U;S_1,V_2)+\epsilon)}$. This also implies that in order to avoid an error the number of words one should use is $R < I(U;Y,S_2,V_2) - I(U;S_1,V_2) - 3\epsilon$, where the last expression also equals $I(U;Y,S_2|V_2) - I(U;S_1|V_2) - 3\epsilon$.

5) As we argued for $P(E_2|E_1^c)$, since $\big(X^n, u^n(1), S_1^n, v_2^n(1)\big)$ is strongly jointly typical, $\big(Y^n, X^n, S_1^n, S_2^n\big)$ is strongly jointly typical and the Markov chain $(U,V_2) - (X,S_1,S_2) - Y$ holds, then, by the Markov lemma [30], $\big(u^n(1), Y^n, S_2^n, v_2^n(1)\big)$ is strongly jointly typical with high probability, i.e., $P(E_5|E_1^c, \ldots, E_4^c) \to 0$.

6)

$$\begin{align}
P(E_6|E_1^c, \ldots, E_5^c) &= \Pr\Big\{ \bigcup_{u^n(l'\neq 1)\in\mathcal{C}_u} \big(u^n(l'), Y^n, S_2^n, v_2^n(1)\big) \in \mathcal{T}_\epsilon^{(n)}(U, Y, S_2, V_2) \Big\} \nonumber\\
&\leq \sum_{l'=2}^{2^{n(I(U;Y,S_2,V_2)+2\epsilon)}} \Pr\Big\{ \big(u^n(l'), Y^n, S_2^n, V_2^n\big) \in \mathcal{T}_\epsilon^{(n)}(U, Y, S_2, V_2) \Big\} \nonumber\\
&\leq \sum_{l'=2}^{2^{n(I(U;Y,S_2,V_2)+2\epsilon)}} 2^{-n(I(U;Y,S_2,V_2)-\epsilon)} \nonumber\\
&\leq 2^{n(I(U;Y,S_2,V_2)-2\epsilon)} 2^{-n(I(U;Y,S_2,V_2)-\epsilon)} \nonumber\\
&= 2^{-n\epsilon}. \tag{67}
\end{align}$$

The probability that there is another index l', $l' \neq 1$, such that $u^n(l')$ is strongly jointly typical with $\big(Y^n, S_2^n, v_2^n(1)\big)$ is bounded by the total number of u^n's times the probability of joint typicality. Therefore, taking $|\mathcal{C}_u| < 2^{n(I(U;Y,S_2,V_2)-\epsilon)}$ assures us that $P(E_6|E_1^c, \ldots, E_5^c) \to 0$. This follows the standard channel capacity argument that one can distinguish at most $2^{nI(U;Y,S_2,V_2)}$ different $u^n(l)$'s given any typical member of $\mathcal{Y}^n \times \mathcal{S}_2^n \times \mathcal{V}_2^n$.

This shows that for rates R and R' as described and for large enough n, the error events are of arbitrarily small probability. This concludes the proof of the achievability and the lower bound on the capacity of Case 2.

Converse: (*Channel capacity Case 2 - Upper bound*). We first prove that it is possible to bound the capacity from above by using two random variables, U and V, that maintain the Markov chain $U - (S_1, V_2) - S_2$ (that is C_2^{ub1}). Then, we prove that it is also possible to upper-bound the capacity by using U and V that maintain the Markov relation $V_2 - S_2 - S_1$ (that is C_2^{ub2}).

Fix the rates R and R' and a sequence of codes $(2^{nR}, 2^{nR'}, n)$ that achieve the capacity. By Fano's inequality, $H(W|Y^n, S_2^n) \leq n\epsilon_n$, where $\epsilon_n \to 0$ as $n \to \infty$. Let $T_2 = f_v(S_2^n)$, and define $V_{2,i} = (T_2, Y^{i-1}, S_{1,i+1}^n, S_2^{i-1})$, $U_i = W$; hence, the Markov chain $U_i - (S_{1,i}, V_{2,i}) - S_{2,i}$ is maintained. The proof for this follows.

$$\begin{aligned}
p(u_i|s_{1,i}, v_{2,i}, s_{2,i}) =& p(w|s_{1,i}, t_2, y^{i-1}, s_{1,i+1}^n, s_2^{i-1}, s_{2,i}) \\
=& \sum_{x^{i-1}, s_1^{i-1}} p(w, x^{i-1}, s_1^{i-1}|s_{1,i}, t_2, y^{i-1}, s_{1,i+1}^n, s_2^{i-1}, s_{2,i}) \\
\overset{(a)}{=}& \sum_{x^{i-1}, s_1^{i-1}} p(s_1^{i-1}|t_2, y^{i-1}, s_{1,i}^n, s_2^{i-1}) p(x^{i-1}|t_2, y^{i-1}, s_1^n, s_2^{i-1}) p(w|x^{i-1}, t_2, y^{i-1}, s_1^n, s_2^{i-1}) \\
=& p(w|t_2, y^{i-1}, s_{1,i+1}^n, s_2^{i-1}, s_{1,i}).
\end{aligned} \tag{68}$$

Next, consider

$$\begin{aligned}
nR' \geq& H(T_2) \\
\geq& H(T_2|S_1^n) - H(T_2|S_1^n, S_2^n) \\
=& I(T_2; S_2^n|S_1^n) \\
=& H(S_2^n|S_1^n) - H(S_2^n|T_2, S_1^n) \\
=& \sum_{i=1}^n \Big[H(S_{2,i}|S_1^n, S_2^{i-1}) - H(S_{2,i}|T_2, S_1^n, S_2^{i-1}) \Big] \\
\overset{(a)}{=}& \sum_{i=1}^n \Big[H(S_{2,i}|S_{1,i}) - H(S_{2,i}|T_2, S_1^n, S_2^{i-1}, Y^{i-1}) \Big] \\
\overset{(b)}{=}& \sum_{i=1}^n \Big[H(S_{2,i}|S_{1,i}) - H(S_{2,i}|T_2, S_{1,i+1}^n, S_2^{i-1}, Y^{i-1}, S_{1,i}) \Big] \\
=& \sum_{i=1}^n \Big[H(S_{2,i}|S_{1,i}) - H(S_{2,i}|V_{2,i}, S_{1,i}) \Big]
\end{aligned}$$

$$=\sum_{i=1}^{n} I(S_{2,i};V_{2,i}|S_{1,i}), \tag{69}$$

where (a) follows from the fact that $S_{2,i}$ is independent of $(S_1^{i-1}, S_{1,i+1}^n, S_2^{i-1})$ given $S_{1,i}$, and the fact that Y^{i-1} is independent of $S_{2,i}$ given (T_2, S_1^n, S_2^{i-1}) (the proof for this follows) and (b) follows from the fact that conditioning reduces entropy.

$$\begin{aligned}
p(y^{i-1}|t_2, s_1^n, s_2^{i-1}, s_{2,i}) &= \sum_{x^n,w} p(y^{i-1}, x^n, w|t_2, s_1^n, s_2^{i-1}, s_{2,i}) \\
&= \sum_{x^n,w} p(w)p(x^n|w, t_2, s_1^n)p(y^{i-1}|x^{i-1}, s_1^{i-1}, s_2^{i-1}) \\
&= p(y^{i-1}|t_2, s_1^n, s_2^{i-1}),
\end{aligned} \tag{70}$$

where we used the facts that W is independent of $(T_2, S_1^n, S_{2,i}^n)$, X^n is a function of (W, T_2, S_1^n) and that the channel is memoryless; i.e., Y^{i-1} is independent of $(W, T_2, S_{1,i}^n, S_{2,i}^n)$ given $(X^{i-1}, S_1^{i-1}, S_2^{i-1})$. We continue the proof of the converse by considering the following set of inequalities:

$$\begin{aligned}
nR =& H(W) \\
\leq& H(W|T_2) - H(W|T_2, Y^n, S_2^n) + n\epsilon_n \\
=& I(W; Y^n, S_2^n|T_2) + n\epsilon_n \\
=& \sum_{i=1}^{n} I(W; Y_i, S_{2,i}|T_2, Y^{i-1}, S_2^{i-1}) + n\epsilon_n \\
\overset{(b)}{=}& \sum_{i=1}^{n} \Big[I(W, S_{1,i+1}^n; Y_i, S_{2,i}|T_2, Y^{i-1}, S_2^{i-1}) \\
& - I(S_{1,i+1}^n; Y_i, S_{2,i}|W, T_2, Y^{i-1}, S_2^{i-1}) \Big] + n\epsilon_n \\
\overset{(c)}{=}& \sum_{i=1}^{n} \Big[I(W, S_{1,i+1}^n; Y_i, S_{2,i}|T_2, Y^{i-1}, S_2^{i-1}) \\
& - I(S_{1,i}; Y^{i-1}, S_2^{i-1}|W, T_2, S_{1,i+1}^n) \Big] + n\epsilon_n \\
=& \sum_{i=1}^{n} \Big[I(W; Y_i, S_{2,i}|T_2, Y^{i-1}, S_{1,i+1}^n, S_2^{i-1}) \\
& - I(S_{1,i}; W|T_2, Y^{i-1}, S_{1,i+1}^n, S_2^{i-1}) \Big] \\
& + \Delta - \Delta^* + n\epsilon_n,
\end{aligned} \tag{71}$$

where

$$\Delta = \sum_{i=1}^{n} I(S_{1,i+1}^n; Y_i, S_{2,i}|T_2, Y^{i-1}, S_2^{i-1}), \tag{72}$$

$$\Delta^* = \sum_{i=1}^{n} I(S_{1,i}; Y^{i-1}, S_2^{i-1}|T_2, S_{1,i+1}^n), \tag{73}$$

(b) follows from the mutual information properties and (c) follows from the Csiszár sum identity.
By using the Csiszár sum on (72) and (73), we get

$$\Delta = \Delta^*, \tag{74}$$

and, therefore, from (79) and (71)

$$\begin{aligned} R' \geq& \frac{1}{n}\sum_{i=1}^{n} I(S_{2,i};V_{2,i}|S_{1,i}) &(75)\\ R-\epsilon_n \leq& \frac{1}{n}\sum_{i=1}^{n}\Big[I(U_i;Y_i,S_{2,i}|V_{2,i})-I(U_i;S_{1,i}|V_{2,i})\Big]. &(76)\end{aligned}$$

Using the convexity of R' and Jansen's inequality, the standard time sharing argument for R and the fact that $\epsilon_n \to 0$ as $n \to \infty$, we can conclude that

$$\begin{aligned} R' \geq& I(V_2;S_2|S_1), &(77)\\ R \leq& I(U;Y,S_2|V_2)-I(U;S_1|V_2), &(78)\end{aligned}$$

where U and V maintain the Markov chain $U-(S_1,V_2)-S_2$.

We now proceed to prove that it is possible to upper-bound the capacity of Case 2 by using two random variables, U and V, that maintain the Markov chain $V_2-S_2-S_1$. Fix the rates R and R' and a sequence of codes $(2^{nR},2^{nR'},n)$ that achieve the capacity. By Fano's inequality, $H(W|Y^n,S_2^n) \leq n\epsilon_n$, where $\epsilon_n \to 0$ as $n \to \infty$. Let $T_2 = f_v(S_2^n)$ and define $V_{2,i}=(T_2,S_2^{i-1}),\ U_i=(W,Y^{i-1},S_{1,i+1}^n)$. The Markov chain $V_{2,i}-S_{2,i}-S_{1,i}$ is maintained. Then,

$$\begin{aligned} nR' \geq& H(T_2)\\ \geq& H(T_2|S_1^n)-H(T_2|S_1^n,S_2^n)\\ =& I(T_2;S_2^n|S_1^n)\\ =& H(S_2^n|S_1^n)-H(S_2^n|T_2,S_1^n)\\ =& \sum_{i=1}^{n}\Big[H(S_{2,i}|S_1^n,S_2^{i-1})-H(S_{2,i}|T_2,S_1^n,S_2^{i-1})\Big]\\ \overset{(a)}{=}& \sum_{i=1}^{n}\Big[H(S_{2,i}|S_{1,i})-H(S_{2,i}|T_2,S_{1,i},S_{1,i+1}^n,S_2^{i-1})\Big]\\ \geq& \sum_{i=1}^{n}\Big[H(S_{2,i}|S_{1,i})-H(S_{2,i}|T_2,S_{1,i},S_2^{i-1})\Big]\\ =& \sum_{i=1}^{n}\Big[H(S_{2,i}|S_{1,i})-H(S_{2,i}|V_{2,i},S_{1,i})\Big]\\ =& \sum_{i=1}^{n} I(S_{2,i};V_{2,i}|S_{1,i}), &(79)\end{aligned}$$

where (a) follows from the fact that $S_{2,i}$ is independent of $(S_1^{i-1}, S_{1,i+1}^n, S_2^{i-1})$ given $S_{1,i}$, and the fact that (Y^{i-1}, S_1^{i-1}) is independent of $S_{2,i}$ given $(T_2, S_{1,i}^n, S_2^{i-1})$; the proof for this follows.

$$\begin{aligned}
p(y^{i-1}, s_1^{i-1}|t_2, s_{1,i}^n, s_2^{i-1}, s_{2,i}) &= \sum_{x^n,w} p(y^{i-1}, s_1^{i-1}, x^n, w|t_2, s_{1,i}^n, s_2^{i-1}, s_{2,i}) \\
&= \sum_{x^n,w} p(w)p(s_1^{i-1}|s_2^{i-1})p(x^n|w, t_2, s_1^n)p(y^{i-1}|x^{i-1}, s_1^{i-1}, s_2^{i-1}) \\
&=p(y^{i-1}, s_1^{i-1}|t_2, s_{1,i}^n, s_2^{i-1}), \qquad (80)
\end{aligned}$$

where we used the facts that W is independent of $(T_2, S_{1,i}^n, S_{2,i}^n)$, S_1^{i-1} is independent of $(T_2, S_{1,i}^n, S_{2,i}^n)$ given S_2^{i-1}, X^n is a function of (W, T_2, S_1^n) and that the channel is memoryless; i.e., Y^{i-1} is independent of $(W, T_2, S_{1,i}^n, S_{2,i}^n)$ given $(X^{i-1}, S_1^{i-1}, S_2^{i-1})$.

In order to complete our proof, we need the following lemma.

Lemma 8. The following inequality holds:

$$\sum_{i=1}^n I(S_{1,i}; W, Y^{i-1}, S_{1,i+1}^n|T_2, S_2^{i-1}) \leq \sum_{i=1}^n I(S_{1,i}; W, Y^{i-1}, S_2^{i-1}|T_2, S_{1,i+1}^n). \qquad (81)$$

Proof: Notice that

$$\sum_{i=1}^n I(S_{1,i}; W, Y^{i-1}, S_{1,i+1}^n|T_2, S_2^{i-1}) = \sum_{i=1}^n I(S_{1,i}; W, Y^{i-1}, S_{1,i+1}^n, S_2^{i-1}|T_2) - I(S_{1,i}; S_2^{i-1}|T_2) \qquad (82)$$

and that

$$\sum_{i=1}^n I(S_{1,i}; W, Y^{i-1}, S_2^{i-1}|T_2, S_{1,i+1}^n) = \sum_{i=1}^n I(S_{1,i}; W, Y^{i-1}, S_{1,i+1}^n, S_2^{i-1}|T_2) - I(S_{1,i}; S_{1,i+1}^n|T_2). \qquad (83)$$

Therefore, it is enough to show that $\sum_{i=1}^n -I(S_{1,i}; S_2^{i-1}|T_2) \leq \sum_{i=1}^n -I(S_{1,i}; S_{1,i+1}^n|T_2)$ holds in order to prove the lemma. Therefore, consider

$$\begin{aligned}
\sum_{i=1}^n -I(S_{1,i}; S_{1,i+1}^n|T_2) - \big(\sum_{i=1}^n -I(S_{1,i}; S_2^{i-1}|T_2)\big) &= \sum_{i=1}^n H(S_{1,i}|T_2, S_{1,i+1}^n) - H(S_{1,i}|T_2, S_2^{i-1}) \\
&= \sum_{i=1}^n H(S_1^n|T_2) - H(S_{1,i}|T_2, S_2^{i-1}) \\
&= \sum_{i=1}^n H(S_{1,i}|T_2, S_1^{i-1}) - H(S_{1,i}|T_2, S_2^{i-1}) \\
&\overset{(a)}{\geq} 0, \qquad (84)
\end{aligned}$$

where (a) follows from the fact that the Markov chain $S_{1,i} - (T_2, S_2^{i-1}) - (T_2, S_1^{i-1})$ holds and from the data processing inequality. This completes the proof of the lemma. ■

We continue the proof of the converse by considering the following set of inequalities:

$$nR =H(W)$$

$$
\begin{aligned}
&\leq H(W|T_2) - H(W|T_2, Y^n, S_2^n) + n\epsilon_n \\
&= I(W; Y^n, S_2^n|T_2) + n\epsilon_n \\
&= \sum_{i=1}^{n} I(W; Y_i, S_{2,i}|T_2, Y^{i-1}, S_2^{i-1}) + n\epsilon_n \\
&\overset{(a)}{=} \sum_{i=1}^{n} \Big[I(W, S_{1,i+1}^n; Y_i, S_{2,i}|T_2, Y^{i-1}, S_2^{i-1}) \\
&\qquad\qquad - I(S_{1,i+1}^n; Y_i, S_{2,i}|W, T_2, Y^{i-1}, S_2^{i-1}) \Big] + n\epsilon_n \\
&\overset{(b)}{=} \sum_{i=1}^{n} \Big[I(W, S_{1,i+1}^n; Y_i, S_{2,i}|T_2, Y^{i-1}, S_2^{i-1}) \\
&\qquad\qquad - I(S_{1,i}; Y^{i-1}, S_2^{i-1}|W, T_2, S_{1,i+1}^n) \Big] + n\epsilon_n \\
&= \sum_{i=1}^{n} \Big[I(W, S_{1,i+1}^n; Y_i, S_{2,i}|T_2, Y^{i-1}, S_2^{i-1}) \\
&\qquad\qquad - I(S_{1,i}; W, Y^{i-1}, S_2^{i-1}|T_2, S_{1,i+1}^n) \Big] + n\epsilon_n \\
&\overset{(c)}{\leq} \sum_{i=1}^{n} \Big[I(W, S_{1,i+1}^n; Y_i, S_{2,i}|T_2, Y^{i-1}, S_2^{i-1}) \\
&\qquad\qquad - I(S_{1,i}; W, Y^{i-1}, S_{1,i+1}^n|T_2, S_2^{i-1}) \Big] + n\epsilon_n \\
&= \sum_{i=1}^{n} I(U_i; Y_i, S_{1,i+1}^n|T_2, S_2^{i-1}) - I(U_i; S_{1,i}|V_{2,i}), \qquad (85)
\end{aligned}
$$

where (a) follows from the mutual information properties, (b) follows from the Csiszár sum identity and (c) follows from Lemma 3. Therefore,

$$
\begin{aligned}
R' &\geq \frac{1}{n} \sum_{i=1}^{n} I(S_{2,i}; V_{2,i}|S_{1,i}) \qquad (86) \\
R - \epsilon_n &\leq \frac{1}{n} \sum_{i=1}^{n} \Big[I(U_i; Y_i, S_{2,i}|V_{2,i}) - I(U_i; S_{1,i}|V_{2,i}) \Big]. \qquad (87)
\end{aligned}
$$

Using the convexity of R' and Jansen's inequality, the standard time sharing argument for R and the fact that $\epsilon_n \to 0$ as $n \to \infty$, we can conclude that

$$
\begin{aligned}
R' &\geq I(V_2; S_2|S_1), \qquad (88) \\
R &\leq I(U; Y, S_2|V_2) - I(U; S_1|V_2), \qquad (89)
\end{aligned}
$$

where the Markov chain $V_2 - S_2 - S_1$ holds. Therefore, we can conclude that the expression given in (12) is an upper-bound to any achievable rate. This concludes the proof of the upper-bound and the proof of Theorem 1 Case 2.

B. Proof of Theorem 1, Case 2_C

For describing the DSI, S_2, with a rate R' we use the standard rate-distortion coding scheme. Then, for the channel coding we use the Shannon strategy [4] coding scheme where the channel's causal state information at the encoder is S_1, S_2 is a part of the channel's output and the rate-limited description of S_2 is the side information at both the encoder and the decoder.

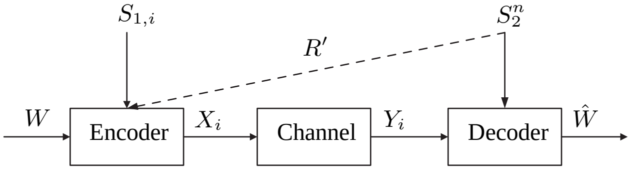

Fig. 16: Channel capacity: Case 2 with causal ESI. $C_{2C} = \max I(U;Y,S_2|V_2)$, where the maximization is over all PMFs $p(v_2|s_2)p(u|v_2)p(x|u,s_1,v_2)$ such that $R' \geq I(V_2;S_2)$.

Achievability: (*Channel capacity Case* 2_C). Given $(S_{1,i}, S_{2,i}) \sim$ i.i.d. $p(s_1,s_2)$, where the ESI is known in a causal way (S_1^i at time i), and the memoryless channel $p(y|x,s_1,s_2)$, fix $p(s_1,s_2,v_2,u,x,y) = p(s_1,s_2)p(v_2|s_2)p(u|v_2)p(x|u,s_1,v_2)p(y|x,s_1,s_2)$, where $x = f(u,s_1,v_2)$ (i.e., $p(x|u,s_1,v_2)$ can get the values 0 or 1).

Codebook generation and random binning

1) Generate a codebook $\mathcal{C}_v$ of $2^{n\left(I(V_2;S_2)+2\epsilon\right)}$ sequences V_2^n independently using i.i.d. $\sim p(v_2)$. Label them $v_2^n(k)$ where $k \in \left\{1,2,\ldots,2^{n(I(V_2;S_2)+2\epsilon)}\right\}$.
2) For each $v_2^n(k)$ generate a codebook $\mathcal{C}_u(k)$ of $2^{n\left(I(U;Y,S_2|V_2)-2\epsilon\right)}$ sequences U^n distributed independently according to i.i.d. $\sim p(u|v_2)$. Label them $u^n(w,k)$, where $w \in \left\{1,2,\ldots,2^{n(I(U;Y,S_2|V_2)-2\epsilon)}\right\}$, and associate the sequences $u^n(w,\cdot)$ with the message $W=w$.

Reveal the codebooks and the content of the bins to all encoders and decoders.

Encoding

1) *State Encoder*: Given the sequence S_2^n, search the codebook $\mathcal{C}_v$ and identify an index k such that $\left(v_2^n(k), S_2^n\right) \in \mathcal{T}_\epsilon^{(n)}(V_2,S_2)$. If such a k is found, stop searching and send it. Otherwise, if no such k is found, declare an error.
2) *Encoder*: Given the message $W \in \left\{1,2,\ldots,2^{n(I(U;Y,S_2|V_2)-2\epsilon)}\right\}$, the index k and S_1^i at time i, identify $u^n(W,k)$ in the codebook $\mathcal{C}_u(k)$ and transmit $x_i = f\left(u_i(W,k), S_{1,i}, v_{2,i}(k)\right)$ at any time $i \in \{1,2,\ldots,n\}$. The element x_i is the result of a multiplexer with an input signal $\left(u_i(W,k), v_{2,i}(k)\right)$ and a control signal $S_{1,i}$.

Decoding

Given Y^n, S_2^n and k, look for a unique index $\hat{W}$, associated with the sequence $u^n(\hat{W},k) \in \mathcal{C}_u(k)$, such that $\left(Y^n, S_2^n, u^n(\hat{W},k)\right) \in \mathcal{T}_\epsilon^{(n)}(Y,U,S_2|v_2^n(k))$. If a unique such $\hat{W}$ is found, declare that the sent message was $\hat{W}$. Otherwise, if no unique index $\hat{W}$ exists, declare an error.

Analysis of the probability of error

Without loss of generality, let us assume that the message $W=1$ was sent and the index k that correspond with

S_2^n is $k=1$; i.e., $v_2^n(1)$ corresponds to S_2^n and $u^n(1,1)$ is chosen according to $\big(W=1, v_2^n(1)\big)$.

Define the following events:

$$\begin{aligned}
E_1 &:= \Big\{\forall v_2^n(k) \in \mathcal{C}_v,\ \big(S_2^n, v_2^n(k)\big) \notin \mathcal{T}_\epsilon^{(n)}(S_2, V_2)\Big\}\\
E_2 &:= \Big\{(u^n(1,1), Y^n, S_2^n) \notin \mathcal{T}_\epsilon^{(n)}(U,Y,S_2|v_2^n(1))\Big\}\\
E_3 &:= \Big\{\exists w' \neq 1:\ u^n(w',1) \in \mathcal{C}_u(1) \text{ and } \big(u^n(w',1), Y^n, S_2^n\big) \in \mathcal{T}_\epsilon^{(n)}(U,Y,S_2|v_2^n(1))\Big\}.
\end{aligned}$$

The probability of error $P_e^{(n)}$ is upper bounded by $P_e^n \leq P(E_1) + P(E_2|E_1^c) + P(E_3|E_1^c, E_2^c)$. Using standard arguments and assuming that $(S_1^n, S_2^n) \in \mathcal{T}_\epsilon^{(n)}(S_1, S_2)$ and that n is large enough, we can state that

1) For each sequence $v_2^n \in \mathcal{C}_v$, the probability that v_2^n is not jointly typical with S_2^n is at most $\big(1 - 2^{-n(I(V_2;S_2)+\epsilon)}\big)$. Therefore, having $2^{n(I(V_2;S_2)+2\epsilon)}$ i.i.d. sequences in $\mathcal{C}_v$, the probability that none of those sequences is jointly typical with S_2^n is bounded by

$$\begin{aligned}
P(E_1) \leq & 2^{n(I(V_2;S_2)+2\epsilon)}\big(1 - 2^{-n(I(V_2;S_2)+\epsilon)}\big)\\
&\leq e^{-2^{n(I(V_2;S_2)+2\epsilon)}2^{-n(I(V_2;S_2)+\epsilon)}}\\
&= e^{-2^{n\epsilon}},
\end{aligned} \tag{90}$$

where, for every $\epsilon > 0$, the last line goes to zero as n goes to infinity.

2) The random variable Y^n is distributed according to $p(y|x,s_1,s_2) = p(y|x,s_1,s_2,v_2)$, therefore, having $(S_2^n, v_2^n(1)) \in \mathcal{T}_\epsilon^{(n)}(S_2,V_2)$ implies that $(Y^n, S_2^n, v_2^n(1)) \in \mathcal{T}_\epsilon^{(n)}(Y,S_2,V_2)$. Recall that $x_i = f\big(u_i(1,1), S_{1,i}, v_2(1)\big)$ and that U^n is generated according to $p(u|v_2)$; therefore, $(X^n, S_1^n, u^n(1,1), v_2^n(1))$ is jointly typical. Thus, by the Markov lemma [30], we can state that $(Y^n, S_2^n, u^n(1,1), v_2^n(1)) \in \mathcal{T}_\epsilon^{(n)}(Y,S_2,U,V_2)$ with high probability for a large enough n.

3) Now, the probability for a random U^n, such that $(U^n, v_2^n(1)) \in \mathcal{T}_\epsilon^{(n)}(U,V_2)$, to be also jointly typical with $(Y^n, S_2^n, v_2^n(1))$ is upper bounded by $2^{-n(I(U,Y,S_2|V_2)-\epsilon)}$, hence

$$\begin{aligned}
P(E_3|E_1^c,E_2^c) &\leq \sum_{1<w'}^{|\mathcal{C}_u(1)|} \Pr\Big\{\big(u^n(w',1), Y^n, S_2^n\big) \in \mathcal{T}_\epsilon^{(n)}(U,Y,S_2|v_2^n(1))\Big\}\\
&\leq \sum_{1<w'}^{|\mathcal{C}_u(1)|} 2^{-n(I(U,Y,S_2|V_2)-\epsilon)}\\
&\leq 2^{n(I(U,Y,S_2|V_2)-2\epsilon)}2^{-n(I(U,Y,S_2|V_2)-\epsilon)}\\
&= 2^{-n\epsilon},
\end{aligned} \tag{91}$$

which goes to zero exponentially fast with n for every $\epsilon > 0$.

Therefore, $P_\epsilon^{(n)} = P(\hat{W} \neq W)$ goes to zero as $n \to \infty$.

Converse: (*Channel capacity case 2*$_c$). Fix the rates R and R' and a sequence of codes $(2^{nR}, 2^{nR'}, n)$ that achieve capacity. By Fano's inequality, $H(W|Y^n, S_2^n) \leq n\epsilon_n$, where $\epsilon_n \to 0$ as $n \to \infty$. Let $T_2 = f_v(S_2^n)$, and define $V_{2,i} = (T_2, Y^{i-1}, S_2^{i-1}),\ U_i = W$. Then,

$$\begin{aligned}
nR' \geq & H(T_2) \\
\geq & H(T_2) - H(T_2|S_2^n) \\
= & I(T_2; S_2^n) \\
= & H(S_2^n) - H(S_2^n|T_2) \\
= & \sum_{i=1}^{n} \Big[H(S_{2,i}|S_2^{i-1}) - H(S_{2,i}|T_2, S_2^{i-1}) \Big] \\
\overset{(a)}{=} & \sum_{i=1}^{n} \Big[H(S_{2,i}) - H(S_{2,i}|T_2, S_2^{i-1}, Y^{i-1}) \Big] \\
= & \sum_{i=1}^{n} I(S_{2,i}; T_2, Y^{i-1}, S_2^{i-1}) \\
= & \sum_{i=1}^{n} I(S_{2,i}; V_{2,i}),
\end{aligned} \tag{92}$$

where (a) follows from the fact that $S_{2,i}$ is independent of S_2^{i-1} and the fact that $S_{2,i}$ is independent of Y^{i-1} given (T_2, S_2^{i-1}). The proof for this follows.

$$\begin{aligned}
p(y^{i-1}|t_2, s_2^{i-1}, s_{2,i}) = & \sum_{w, x^{i-1}, s_1^{i-1}} p(y^{i-1}, w, x^{i-1}, s_1^{i-1}|t_2, s_2^{i-1}, s_{2,i}) \\
= & \sum_{w, x^{i-1}, s_1^{i-1}} p(w) p(s_1^{i-1}|s_2^{i-1}) p(x^{i-1}|w, t_2, s_1^{i-1}) p(y^{i-1}|x^{i-1}, s_1^{i-1}, s_2^{i-1}) \\
= & p(y^{i-1}|t_2, s_2^{i-1}),
\end{aligned} \tag{93}$$

where we used the fact that W is independent of $(T_2, S_2^{i-1}, S_{2,i})$, S_1^{i-1} is independent of $(T_2, S_{2,i})$ given S_2^{i-1}, X^{i-1} is a function of (W, T_2, S_1^{i-1}) and that Y^{i-1} is independent of $(W, T_2, S_{2,i})$ given $(X^{i-1}, S_1^{i-1}, S_2^{i-1})$. We now continue with the proof of the converse.

$$\begin{aligned}
nR \leq & H(W) \\
\leq & H(W|T_2) - H(W|T_2, Y^n, S_2^n) + n\epsilon_n \\
= & I(W; Y^n, S_2^n|T_2) + n\epsilon_n \\
= & \sum_{i=1}^{n} I(W; Y_i, S_{2,i}|T_2, Y^{i-1}, S_2^{i-1}) + n\epsilon_n \\
= & \sum_{i=1}^{n} I(U_i; Y_i, S_{2,i}|V_{2,i}) + n\epsilon_n
\end{aligned} \tag{94}$$

and therefore, from (92) and (94)

$$R' \geq \frac{1}{n} \sum_{i=1}^{n} I(S_{2,i}; V_{2,i}) \tag{95}$$

$$R-\epsilon_n \leq \frac{1}{n}\sum_{i=1}^{n} I(U_i;Y_i,S_{2,i}|V_{2,i}). \tag{96}$$

Using the convexity of R' and Jansen's inequality, the standard time-sharing argument for R and the fact that $\epsilon_n \to 0$ as $n \to \infty$, we can conclude that

$$R' \geq I(V_2;S_2), \tag{97}$$
$$R \leq I(U;Y,S_2|V_2). \tag{98}$$

Notice that the Markov chain $V_{2,i} - S_{2,i} - S_{1,i}$ holds since (Y^{i-1},S_2^{i-1}) is independent of $S_{1,i}$ and $T_2(S_2^n)$ is dependent on $S_{1,i}$ only through $S_{2,i}$. Notice also that the Markov chain $U_i - V_{2,i} - (S_{1,i},S_{2,i})$ holds since

$$\begin{aligned}
p(w|t_2,y^{i-1},s_2^{i-1},s_{1,i},s_{2,i}) &= \sum_{x^{i-1},s_1^{i-1}} p(w,x^{i-1},s_1^{i-1}|t_2,y^{i-1},s_2^{i-1},s_{1,i},s_{2,i}) \\
&= \sum_{x^{i-1},s_1^{i-1}} p(s_1^{i-1}|t_2,y^{i-1},s_2^{i-1})p(x^{i-1}|t_2,y^{i-1},s_1^{i-1},s_2^{i-1})p(w|t_2,x^{i-1},s_1^{i-1}) \\
&= p(w|t_2,y^{i-1},s_2^{i-1}).
\end{aligned} \tag{99}$$

This concludes the converse, and the proof of Theorem 1 Case 2_C.

APPENDIX C
PROOF OF THEOREM 2

In this section we provide the proof of Theorem 2, Cases 1 and 1_C. Case 2, where the encoder is informed with increased ESI and the decoder is informed with DSI is a special case of [10] for $K=1$ and, therefore, the proof for this case is omitted. Following Kaspi's scheme (Figure 17) for $K=1$, at the first stage, node W sends a description of W with a rate limited to R_w, then, after reconstructing $\hat{W}$ at the Z node, it sends a function of Z and $\hat{W}$ over to node W with a rate limited to R_z. Let S_2 be W in Kaspi's scheme and (X,S_1) be Z in Kaspi's scheme. Consider $D_z = d(Z_i,\hat{Z}_i) = d\big((X,S_{1,i}),(\hat{X}_i,\hat{S}_{1,i})\big) = d(X_i,\hat{X}_i) = D$. Then, it is apparent that Case 2 of the rate-distortion problems is a special case of Kaspi's two-way problem for $K=1$.

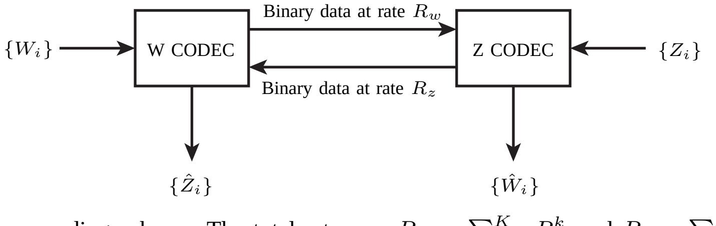

Fig. 17: Kaspi's two-way source coding scheme. The total rates are $R_w = \sum_{k=1}^{K} R_w^k$ and $R_z = \sum_{k=1}^{K} R_z^k$ and the expected per-letter distortions are $D_w = \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} d(W_i,\hat{W}_i)\Big]$ and $D_z = \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} d(Z_i,\hat{Z}_i)\Big]$.

A. Proof of Theorem 2, Case 1

We use the Wyner-Ziv coding scheme for the description of the ESI, S_1, at a rate R', where the source is S_1 and the side information at the decoder is S_2. Then, to describe the main source, X, with distortion less than or equal to D we use the Wyner-Ziv coding scheme again, where this time, S_2 is the side information at the decoder, S_1 is a part of the source and the rate-limited description of S_1 is the side information at both the encoder and the decoder. Notice that $I(U;X,S_1|V_1)-I(U;S_2|V_1)=I(U;X,S_1,V_1)-I(U;S_1,V_1)$ and that since the Markov chain $V_1-S_1-S_2$ holds, it is also possible to write $R'\geq I(V_1;S_1)-I(V_1;S_2)$; we use these expressions in the following proof.

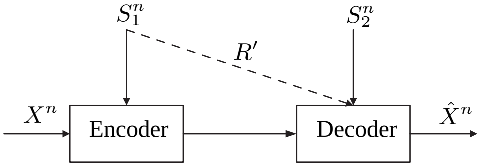

Fig. 18: Rate-distortion: Case 1. $R_1(D)=\min I(U;X,S_1|V_1)-I(U;S_2|V_1)$, where the minimization is over all PMFs $p(v_1|s_1)p(u|x,s_1,v_1)p(\hat{x}|u,s_2,v_1)$ such that $R'\geq I(V_1;S_1|S_2)$ and $\mathbb{E}\Big[d(X,\hat{X})\Big]\leq D$.

Achievability: (*Rate-distortion Case 1*). Given $(X_i,S_{1,i},S_{2,i})$ i.i.d. $\sim$ $p(x,s_1,s_2)$ and the distortion measure D, fix $p(x,s_1,s_2,v_1,u,\hat{x})=p(x,s_1,s_2)p(v_1|s_1)p(u|x,s_1,v_1)p(\hat{x}|u,s_2,v_1)$ that satisfies $\mathbb{E}\big[d(X,\hat{X})\big]=D$ and $\hat{x}=f(u,s_2,v_1)$.

Codebook generation and random binning

1) Generate a codebook, $\mathcal{C}_v$, of $2^{n\big(I(V_1;S_1)+2\epsilon\big)}$ sequences, V_1^n, independently using i.i.d. $\sim p(v_1)$. Label them $v_1^n(k)$, where $k\in\big\{1,2,\dots,2^{n(I(V_1;S_1)+2\epsilon)}\big\}$ and randomly assign each sequence $v_1^n(k)$ a bin number $b_v\big(v_1^n(k)\big)$ in the set $\big\{1,2,\dots,2^{nR'}\big\}$.
2) Generate a codebook $\mathcal{C}_u$ of $2^{n\big(I(U;X,S_1,V_1)+2\epsilon\big)}$ sequences U^n independently using i.i.d. $\sim p(u)$. Label them $u^n(l)$, where $l\in\big\{1,2,\dots,2^{n(I(U;X,S_1,V_1)+2\epsilon)}\big\}$, and randomly and assign each $u^n(l)$ a bin number $b_u\big(u^n(l)\big)$ in the set $\big\{1,2,\dots,2^{nR}\big\}$.

Reveal the codebooks and the content of the bins to all encoders and decoders.

Encoding

1) *State Encoder*: Given the sequence S_1^n, search the codebook $\mathcal{C}_v$ and identify an index k such that $\big(S_1^n,v_1^n(k)\big)\in\mathcal{T}_\epsilon^{(n)}(S,V_1)$. If such a k is found, stop searching and send the bin number $j=b_v\big(v_1^n(k)\big)$. If no such k is found, declare an error.
2) *Encoder*: Given the sequences X^n, S_1^n and $v_1^n(k)$, search the codebook $\mathcal{C}_u$ and identify an index l such that $\big(X^n,S_1^n,v_1^n(k),u^n(l)\big)\in\mathcal{T}_\epsilon^{(n)}(X,S_1,V_1,U)$. If such an l is found, stop searching and send the bin number $w=b_u\big(u^n(l)\big)$. If no such l is found, declare an error.

Decoding

Given the bins indices w and j and the sequence S_2^n, search the codebook $\mathcal{C}_v$ and identify an index k such that $\big(S_2^n, v_1^n(k)\big) \in \mathcal{T}_\epsilon^{(n)}(S_2, V_1)$ and $b_v\big(v_1^n(k)\big) = j$. If no such k is found or there is more than one such index, declare an error. If a unique k, as defined, is found, search the codebook $\mathcal{C}_u$ and identify an index l such that $\big(S_2^n, v_1^n(k), u^n(l)\big) \in \mathcal{T}_\epsilon^{(n)}(S_2, V_1, U)$ and $b_u\big(u^n(l)\big) = w$. If a unique l, as defined, is found, declare $\hat{X}_i = f_i(u_i^n(l), S_{2,i}, v_{1,i}(k)),\ i = 1, 2, \ldots, n$. Otherwise, if there is no such l or there is more than one, declare an error.

Analysis of the probability of error

Without loss of generality, for the following events E_2, E_3, E_4, E_5 and E_6, assume that $v_1^n(k=1)$ and $b_v\big(v_1^n(k = 1)\big) = 1$ correspond to the sequences (X^n, S_1^n, S_2^n) and for the events E_5 and E_6 assume that $u^n(l = 1)$ and $b_u\big(u^n(l = 1)\big) = 1$ correspond to the same given sequences. Define the following events:

$$
\begin{aligned}
E_1 &:= \Big\{\forall v_1^n(k) \in \mathcal{C}_v,\ \big(S_1^n, v_1^n(k)\big) \notin \mathcal{T}_\epsilon^{(n)}(S_1, V_1)\Big\} \\
E_2 &:= \Big\{\big(S_1^n, v_1^n(1)\big) \in \mathcal{T}_\epsilon^{(n)}(S_1, V_1) \text{ but } \big(S_2^n, v_1^n(1)\big) \notin \mathcal{T}_\epsilon^{(n)}(S_2, V_1)\Big\} \\
E_3 &:= \Big\{\exists k' \neq 1 \text{ such that } b_v\big(v_1^n(k')\big) = 1 \text{ and } \big(S_2^n, v_1^n(k')\big) \in \mathcal{T}_\epsilon^{(n)}(S_2, V_1)\Big\} \\
E_4 &:= \Big\{\forall u^n(l) \in \mathcal{C}_u,\ \big(X^n, S_1^n, v_1^n(1), u^n(l)\big) \notin \mathcal{T}_\epsilon^{(n)}(X, S_1, V_1, U)\Big\} \\
E_5 &:= \Big\{\big(X^n, S_1^n, v_1^n(1), u^n(1)\big) \in \mathcal{T}_\epsilon^{(n)}(X, S_1, V_1, U\big) \text{ but } \big(S_2^n, v_1^n(1), u^n(1)\big) \notin \mathcal{T}_\epsilon^{(n)}(S_2, V_1, U\big)\Big\} \\
E_6 &:= \Big\{\exists l' \neq 1 \text{ such that } b_u\big(u^n(l')\big) = 1 \text{ and } \big(S_2^n, v_1^n(1), u^n(l')\big) \in \mathcal{T}_\epsilon^{(n)}(S_2, V_1, U)\Big\}.
\end{aligned}
$$

The probability of error $P_e^{(n)}$ is upper bounded by $P_e^n \leq P(E_1) + P(E_2|E_1^c) + P(E_3|E_1^c, E_2^c) + P(E_4|E_1^c, E_2^c, E_3^c) + P(E_5|E_1^c, \ldots, E_4^c) + P(E_6|E_1^c \ldots, E_5^c)$. Using standard arguments and assuming that $(X^n, S_1^n, S_2^n) \in \mathcal{T}_\epsilon^{(n)}(X, S_1, S_2)$ and that n is large enough, we can state that

1)

$$
\begin{aligned}
P(E_1) =& \Pr\Big\{ \bigcap_{v_1^n(k) \in \mathcal{C}_v} \big(S_1^n, v_1^n(k)\big) \notin \mathcal{T}_\epsilon^{(n)}(S_1, V_1)\Big\} \\
\leq& \prod_{k=1}^{2^{n\left(I(V_1;S_1)+\epsilon\right)}} \Pr\{\big(S_1^n, V_1^n(k)\big) \notin \mathcal{T}_\epsilon^{(n)}(S_1, V_1)\} \\
\leq& e^{-2^{n\left(I(V_1;S_1)+2\epsilon\right)} 2^{-nI(S_1;V_1)-n\epsilon}} \\
=& e^{-n\epsilon}. \qquad (100)
\end{aligned}
$$

The probability that there is no $v_1^n(k)$ in $\mathcal{C}_v$ such that $\big(S_1^n, v_1^n(k)\big)$ is strongly jointly typical is exponentially small provided that $|\mathcal{C}_v| > 2^{n\left(I(S_1;V_1)+\epsilon\right)}$. This follows from the standard rate-distortion argument that $2^{nI(S_1;V_1)}$ $v_1^n(k)$s “cover” $\mathcal{S}_1^n$, therefore $P(E_1) \to 0$.

2) By the Markov lemma, since (S_1^n, S_2^n) are strongly jointly typical and $\big(S_1^n, v_1^n(1)\big)$ are strongly jointly typical and the Markov chain $V_1 - S_1 - S_2$ holds, then $\big(S_1^n, S_2^n, v_1^n(1)\big)$ are also strongly jointly typical. Thus, $P(E_2|E_1^c) \to 0$.

3)

$$\begin{aligned}P(E_3) =&\Pr\Big\{ \bigcup_{\substack{v_1^n(k'\neq 1)\\ b_v\big(v_1(k')\big)=1}} \big(S_2^n, v_1^n(k')\big) \in \mathcal{T}_\epsilon^{(n)}(S_1,V_1)\Big\}\\ \leq& \sum_{\substack{v_1^n(k'\neq 1)\\ b_v\big(v_1(k')\big)=1}} \Pr\big\{\big(S_1^n, v_1^n(k')\big) \in \mathcal{T}_\epsilon^{(n)}(S_1,V_1)\big\}\\ \leq& 2^{n\big(I(V_1;S_1)+2\epsilon-R'\big)}2^{-n\big(I(S_2;V_1)-\epsilon\big)}. \end{aligned} \tag{101}$$

The probability that there is another index $k',\ k'\neq 1$, such that $v_1^n(k')$ is in bin number 1 and that it is strongly jointly typical with S_2^n is bounded by the number of $v_1^n(k)$'s in the bin times the probability of joint typicality. Therefore, if $R' > I(V_1;S_1) - I(V_1;S_2) + 3\epsilon$ then $P(E_3|E_1^c,E_2^c) \to 0$. Furthermore, using the Markov chain $V_1 - S_1 - S_2$, we can see that the inequality can be presented as $R' > I(V_1;S_1|S_2) + 3\epsilon$.

4) We use here the same argument we used for $P(E_1)$. By the covering lemma we can state that the probability that there is no $u^n(l)$ in $\mathcal{C}_u$ that is strongly jointly typical with $\big(X^n, S_1^n, v_1^n(k)\big)$ tends to 0 as $n\to\infty$ if $R'_u > I(U;X,S_1,V_1)+\epsilon$. Hence, $P(E_4|E_1^c,E_2^c,E_3^c)\to 0$.
5) Using the same argument we used for $P(E_2|E_1^c)$, we conclude that $P(E_4|E_1^c,E_2^c,E_3^c)\to 0$.
6) We use here the same argument we used for $P(E_2|E_1^c)$. Since (U,X,S_1V_1) are strongly jointly typical, (X,S_1,S_2) are strongly jointly typical and the Markov chain $(U,V_1)-(X,S_1)-S_2$ holds, then (U,X,S_1,S_2,V_1) are also strongly jointly typical.
7) The probability that there is another index $l',\ l'\neq 1$ such that $u^n(l')$ is in bin number 1 and that it is strongly jointly typical with $\big(S_2^n, v_1^n(1)\big)$ is exponentially small provided that $R \geq I(U;X,S_1,V_1) - I(U;S_2,V_1) + 3\epsilon = I(U;X,S_1|V_1) - I(U;S_2|V_1) + 3\epsilon$. Notice that $2^{n(I(U;X,S_1,V_1)-R)}$ stands for the average number of sequences $u^n(l)$'s in each bin indexed w for $w\in\{1,2,\ldots,2^{nR}\}$.

This shows that for rates R and R' as described, and for large enough n, the error events are of arbitrarily small probability. This concludes the proof of the achievability for the source coding Case 1.

Converse: (*Rate-distortion Case 1*). Fix a distortion measure D, the rates R', $R \geq R(D) = \min I(U;X,S_1|V_1) - I(U;S_2|V_1) = \min I(U;X,S_1|S_2,V_1)$ and a sequence of codes $(2^{nR}, 2^{nR'}, n)$ such that $\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n d(X_i,\hat{X}_i)\Big] = D$. Let $T_1 = f_v(S_1^n)$, $T = f(X^n, S_1^n, T)$ and define $V_{1,i} = (T_1, S_{1,i+1}^n, S_2^{i-1}, S_{2,i+1}^n)$ and $U_i = T$. Notice that $\hat{X}_i = \hat{X}_i(T,T_1,S_2^n)$ and, therefore, $\hat{X}_i$ is a function of $(U_i, V_{1,i}, S_{2,i})$.

$$\begin{aligned} nR' \geq& H(T_1)\\ \geq& H(T_1|S_2^n) - H(T_1|S_1^n,S_2^n)\\ =& I(T_1;S_1^n|S_2^n)\\ =& H(S_1^n|S_2^n) - H(S_1^n|T_1,S_2^n)\\ =& \sum_{i=1}^n \Big[H(S_{1,i}|S_{1,i+1}^n,S_2^n) - H(S_{1,i}|T_1,S_{1,i+1}^n,S_2^n)\Big] \end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{=}\sum_{i=1}^{n}\Big[H(S_{1,i}|S_{2,i})-H(S_{1,i}|T_1,S_{1,i+1}^{n},S_2^{i-1},S_{2,i+1}^{n},S_{2,i})\Big]\\
&=\sum_{i=1}^{n}\Big[H(S_{1,i}|S_{2,i})-H(S_{1,i}|V_{1,i},S_{2,i})\Big]\\
&=\sum_{i=1}^{n}I(S_{1,i};V_{1,i}|S_{2,i}),
\end{aligned}
\tag{102}
$$

where (a) follows from the fact that $S_{1,i}$ is independent of $(S_{1,i+1}^{n},S_2^{i-1},S_{2,1+i}^{n})$ given $S_{2,i}$.

$$
\begin{aligned}
nR\geq&H(T)\\
\geq&H(T|T_1,S_2^n)-H(T|T_1,X^n,S_1^n,S_2^n)\\
=&I(T;X^n,S_1^n|T_1,S_2^n)\\
=&H(X^n,S_1^n|T_1,S_2^n)-H(X^n,S_1^n|T,T_1,S_2^n)\\
=&\sum_{i=1}^{n}\Big[H(X_i,S_{1,i}|T_1,S_2^n,X_{i+1}^n,S_{1,i+1}^n)-H(X_i,S_{1,i}|T,T_1,S_2^n,X_{i+1}^n,S_{1,i+1}^n)\Big]\\
\overset{(b)}{=}&\sum_{i=1}^{n}\Big[H(X_i,S_{1,i}|T_1,S_{1,i+1}^n,S_2^n)-H(X_i,S_{1,i}|T,T_1,S_2^n,X_{i+1}^n,S_{1,i+1}^n)\Big]\\
\overset{(c)}{\geq}&\sum_{i=1}^{n}\Big[H(X_i,S_{1,i}|T_1,S_{1,i+1}^n,S_2^n)-H(X_i,S_{1,i}|T,T_1,S_{1,i+1}^n,S_2^n)\Big]\\
=&\sum_{i=1}^{n}I(X_i,S_{1,i};T|T_1,S_{1,i+1}^n,S_2^n)\\
=&\sum_{i=1}^{n}I(X_i,S_{1,i};U_i|V_{1,i},S_{2,i})\\
=&\sum_{i=1}^{n}R\Big(\mathbb{E}\Big[d\big(X_i,\hat{X}_i\big)\Big]\Big)\\
\overset{(d)}{\geq}&nR\Big(\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}d\big(X_i,\hat{X}_i\big)\Big]\Big)\\
=&nR(D),
\end{aligned}
\tag{103}
$$

where (b) follows from the fact that $(X_i,S_{1,i})$ is independent of X_{i+1}^n given $(T_1,S_{1,i+1}^n,S_2^n)$; this is because X_{i+1}^n is independent of (T_1,X^i,S_1^i) given $(S_{1,i+1}^n,S_{2,i+1}^n)$, (c) follows from the fact that conditioning reduces entropy and (d) follows from the convexity of $R(D)$ and Jensen's inequality.

Using also the convexity of R' and Jensen's inequality, we can conclude that

$$R'\geq I(V_1;S_1|S_2), \tag{104}$$

$$R\geq I(U;X,S_1|V_1,S_2). \tag{105}$$

It is easy to verify that $(T_1,S_{1,i+1}^n,S_2^{i-1},S_{2,i+1}^n)-S_{1,i}-S_{2,i}$ forms a Markov chain, since $T_1(S_1^n)$ depends on $S_{2,i}$ only through $S_{1,i}$. The structure $T-(T_1,S_{1,i+1}^n,S_2^{i-1},S_{2,i+1}^n,X_i,S_{1,i})-S_{2,i}$ also forms a Markov chain since $S_{2,i}$ contains no information about $(S_1^{i-1},X^{i-1},X_{i+1}^n)$ given $(T_1,S_{1,i}^n,S_2^{i-1},S_{2,i+1}^n,X_i)$ and, therefore, contains no information about $T(X^n, S_1^n, T_1)$.

This concludes the converse, and the proof of Theorem 2 Case 1.

B. Proof of Theorem 2, Case 1_C

For describing the ESI, S_1, with a rate R' we use the standard rate-distortion coding scheme. Then, for the main source, X, we use a Weissman-El Gamal [12] coding scheme where the DSI, S_2, is the causal side information at the decoder, S_1 is a part of the source and the rate-limited description of S_1 is the side information at both the encoder and decoder.

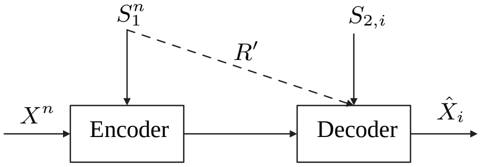

Fig. 19: Rate-distortion: Case 1 with causal DSI. $R_{1C}(D) = \min I(U; X, S_1|V_1)$, where the minimization is over all PMFs $p(v_1|s_1)p(u|x, s_1, v_1)p(\hat{x}|u, s_2, v_1)$ such that $R' \geq I(V_1; S_1)$ and $\mathbb{E}\left[d(X, \hat{X})\right] \leq D$.

Achievability: (*Rate-distortion Case 1_C*). Given $(X_i, S_{1,i}, S_{2,i}) \sim$ i.i.d. $p(x, s_1, s_2)$ where the DSI is known in a causal way (S_2^i in time i) and the distortion measure is D, fix $p(x, s_1, s_2, v_1, u, \hat{x}) = p(x, s_1, s_2)p(v_1|s_1)p(u|x, s_1, v_1)p(\hat{x}|u, s_2, v_1)$ that satisfies $\mathbb{E}\left[d(X, \hat{X})\right] = D$ and that $\hat{x} = f(u, s_2, v_1)$.

Codebook generation and random binning

1) Generate a codebook $\mathcal{C}_v$ of $2^{n\left(I(V_1;S_1)+2\epsilon\right)}$ sequences V_1^n independently using i.i.d. $\sim p(v_2)$. Label them $v_1^n(k)$ where $k \in \left\{1, 2, \ldots, 2^{n(I(V_1;S_1)+2\epsilon)}\right\}$.
2) For each $v_1^n(k)$ generate a codebook $\mathcal{C}_u(k)$ of $2^{n\left(I(U;X,S_1|V_1)+2\epsilon\right)}$ sequences U^n distributed independently according to i.i.d. $\sim p(u|v_1)$. Label them $u^n(w, k)$, where $w \in \left\{1, 2, \ldots, 2^{n(I(U;X,S_1|V_1)+2\epsilon)}\right\}$.

Reveal the codebooks to all encoders and decoders.

Encoding

1) *State Encoder*: Given the sequence S_1^n, search the codebook $\mathcal{C}_v$ and identify an index k such that $\left(v_1^n(k), S_1^n\right) \in \mathcal{T}_\epsilon^{(n)}(V_1, S_1)$. If such a k is found, stop searching and send it. Otherwise, if no such k is found, declare an error.
2) *Encoder*: Given X^n, S_1^n and the index k, search the codebook $\mathcal{C}_u(k)$ and identify an index w such that $\left(u^n(w, k), X^n, S_1^n\right) \in \mathcal{T}_\epsilon^{(n)}(U, X, S_1|v_1^n(k))$. If such an index w is found, stop searching and send it. Otherwise, declare an error.

Decoding

Given the indices w, k and the sequence S_1^i at time i, declare $\hat{x}_i = f\left(u_i(w, k), S_{2,i}, v_{1,i}(k)\right)$.

Analysis of the probability of error

Without loss of generality, let us assume that $v_1^n(1)$ corresponds to S_1^n and that $u^n(1, 1)$ corresponds to

$(X^n, S_1^n, v_1^n(1))$.

Define the following events:

$$E_1 := \left\{\forall v_1^n(k) \in \mathcal{C}_v,\ \left(v_1^n(k), S_1^n\right) \notin \mathcal{T}_\epsilon^{(n)}(S_1, V_1)\right\}$$
$$E_2 := \left\{\forall u^n(w,1) \in \mathcal{C}_u(1),\ \left(X^n, S_1^n, u^n(w,1)\right) \notin \mathcal{T}_\epsilon^{(n)}(X, S_1, U)\right\}$$

The probability of error $P_e^{(n)}$ is upper bounded by $P_e^n \leq P(E_1) + P(E_2|E_1^c)$. Assuming that $(S_1^n, S_2^n) \in \mathcal{T}_\epsilon^{(n)}(S_1, S_2)$, we can state that by the standard rate-distortion argument, having more than $2^{n(I(V_1;S_1)+\epsilon)}$ sequences $v_1^n(k)$ in $\mathcal{C}_v$ and a large enough n assures us with probability arbitrarily close to 1 that we would find an index k such that $\left(v_1^n(k), S_1^n\right) \in \mathcal{T}_\epsilon^{(n)}(V_1, S_1)$. Therefore, $P(E_1) \to 0$ as $n \to \infty$. Now, if $\left(v_1^n(1), S_1^n\right) \in \mathcal{T}_\epsilon^{(n)}(V_1, S_1)$, using the same argument, we can also state that having more than $2^{n(I(U;X,S_1|V_1)+\epsilon)}$ sequences $u^n(w,1)$ in $\mathcal{C}_u(1)$ assures us that $P(E_2|E_1^c) \to 0$ as $n \to \infty$. This concludes the proof of the achievability.

Converse: (*Rate-distortion Case 1_C*). Fix a distortion measure D, the rates R', $R \geq R(D) = \min I(U; X, S_1|V_1)$ and a sequence of codes $(2^{nR}, 2^{nR'}, n)$ such that $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n d(X_i, \hat{X}_i)\right] = D$. Let $T_1 = f_v(S_1^n)$, $T = f(X^n, S_1^n, T_1)$ and define $V_{1,i} = (T_1, S_{1,i+1}^n)$, $U_i = T$. Notice that $\hat{X}_i = \hat{X}_i(T, T_1, S_2^i)$, and, therefore, $\hat{X}_i$ is a function of $(U_i, V_{1,i}, S_2^i)$.

$$\begin{aligned}
nR' \geq & H(T_1) \\
\geq & H(V) - H(T_1|S_1^n) \\
= & I(T_1; S_1^n) \\
= & H(S_1^n) - H(S_1^n|T_1) \\
= & \sum_{i=1}^n \left[H(S_{1,i}|S_{1,i+1}^n) - H(S_{1,i}|T_1, S_{1,i+1}^n)\right] \\
\overset{(a)}{=} & \sum_{i=1}^n \left[H(S_{1,i}) - H(S_{1,i}|T_1, S_{1,i+1}^n)\right] \\
= & \sum_{i=1}^n \left[H(S_{1,i}) - H(S_{1,i}|V_{1,i})\right] \\
= & \sum_{i=1}^n I(S_{1,i}; V_{1,i}), \qquad (106)
\end{aligned}$$

where (a) follows the fact that $S_{1,i}$ is independent of $S_{1,i+1}^n$.

$$\begin{aligned}
nR \geq & H(T) \\
\geq & H(T|T_1) - H(T|T_1, X^n, S_1^n) \\
= & I(T; X^n, S_1^n|T_1) \\
= & H(X^n, S_1^n|T_1) - H(X^n, S_1^n|T, T_1) \\
= & \sum_{i=1}^n \left[H(X_i, S_{1,i}|T_1, X_{i+1}^n, S_{1,i+1}^n) - H(X_i, S_{1,i}|T, T_1, X_{i+1}^n, S_{1,i+1}^n)\right]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{=}\sum_{i=1}^{n}\Big[H(X_i,S_{1,i}|T_1,S_{1,i+1}^n)-H(X_i,S_{1,i}|T,T_1,X_{i+1}^n,S_{1,i+1}^n)\Big]\\
&\overset{(c)}{\geq}\sum_{i=1}^{n}\Big[H(X_i,S_{1,i}|T_1,S_{1,i+1}^n)-H(X_i,S_{1,i}|T,T_1,S_{1,i+1}^n)\Big]\\
&=\sum_{i=1}^{n}I(X_i,S_{1,i};T|T_1,S_{1,i+1}^n)\\
&=\sum_{i=1}^{n}I(X_i,S_{1,i};U_i|V_{1,i})\\
&=\sum_{i=1}^{n}R\Big(\mathbb{E}\Big[d\big(X_i,\hat{X}_i\big)\Big]\Big)\\
&\overset{(d)}{\geq}nR\Big(\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}d\big(X_i,\hat{X}_i\big)\Big]\Big)\\
&=nR(D)
\end{aligned}
\tag{107}
$$

where (b) follows from the fact that $(X_i,S_{1,i})$ is independent of X_{i+1}^n given $(T_1,S_{1,i+1}^n)$, (c) follows from the fact that conditioning reduces entropy and (d) follows from the convexity of $R(D)$ and Jensen's inequality.

Using also the convexity of R' and Jensen's inequality, we can conclude that

$$R' \geq I(V_1;S_1), \tag{108}$$

$$R \geq I(U;X,S_1|V_1). \tag{109}$$

It is easy to verify that both Markov chains $V_{1,i}-S_{1,i}-(X_i,S_{2,i})$ and $U_i-(X_i,S_{1,i},V_{1,i})-S_{2,i}$ hold. This concludes the converse, and the proof of Theorem 2 Case 1_C.

C. Proof of Theorem 2, Case 2

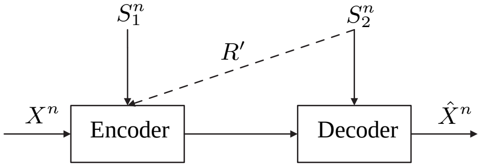

Fig. 20: Rate distortion: Case 2. $R_2(D)=\min I(U;X,S_1|V_2)-I(U;S_2|V_2)$, where the minimization is over all PMFs $p(v_2|s_2)p(u|x,s_1,v_2)p(\hat{x}|u,s_2,v_2)$ such that $R'\geq I(V_2;S_2)-I(V_2;X,S_1)$ and $\mathbb{E}\Big[d(X,\hat{X})\Big]\leq D$.

This problem is a special case of [10] for $K=1$, and hence, the proof is omitted.

Appendix D
Proof of Lemma 1

We provide here a partial proof of Lemma 1. In the first part we prove the concavity of $C_2^{lb}(R')$ in R' for Case 2, the second part contains the proof that it is enough to take X to be a deterministic function of (S_1,V_1,U) in order to achieve the capacity $C_1(R')$ for Case 1 and in the third part we prove the cardinality bound for Case 1. The proofs of these three parts for the rest of the cases can be derived using the same techniques and therefore are omitted. The proof of Lemma 2 can also be readily concluded using the techniques we use in this appendix and is omitted as well.

Part 1: We prove here that for Case 2 of the channel capacity problems, the lower bound on the capacity, $C_2^{lb}(R')$, is a concave function of the state information rate, R'. Recall that the expression for C_2^{lb} is $C_2^{lb}(R') = \max I(U;Y,S_2|V_2) - I(U;S_1|V_2)$ where the maximization is over all probabilities $p(s_1,s_2)p(v_2|s_2)p(u|s_1,v_2)p(x|u,s_1,v_2)p(y|x,s_1,s_2)$ such that $R' \geq I(V_2;S_2|S_1)$. This means that we want to prove that for any two rates, $R'^{(1)}$ and $R'^{(2)}$, and for any $0 \leq \alpha \leq 1$ and $\bar{\alpha} = 1-\alpha$ the capacity maintains $C_2^{lb}\big(\alpha R'^{(1)} + \bar{\alpha}R'^{(2)}\big) \geq \alpha C_2^{lb}(R'^{(1)}) + \bar{\alpha}C_2^{lb}(R'^{(2)})$. Let $(U^{(1)},V_2^{(1)},X^{(1)},Y^{(1)})$ and $(U^{(2)},V_2^{(2)},X^{(2)},Y^{(2)})$ be the random variables that meet the conditions on $R'^{(1)}$ and on $R'^{(2)}$ and also achieve $C_2^{lb}(R'^{(1)})$ and $C_2^{lb}(R'^{(2)})$, respectively. Let us introduce the auxiliary random variable $Q \in \{1,2\}$, independent of S_1,S_2,V_2,U,X and Y, and distributed according to $\Pr\{Q=1\} = \alpha$ and $\Pr\{Q=2\} = \bar{\alpha}$. Then, consider

$$\begin{aligned}\alpha R'^{(1)} + \bar{\alpha}R'^{(2)} &= \alpha\big[I(V_2^{(1)};S_2) - I(V_2^{(1)};S_1)\big] + \bar{\alpha}\big[I(V_2^{(2)};S_2) - I(V_2^{(2)};S_1)\big]\\ &\overset{(a)}{=} \alpha\big[I(V_2^{(1)};S_2|Q=1) - I(V_2^{(1)};S_1|Q=1)\big] + \bar{\alpha}\big(I(V_2^{(2)};S_2|Q=2) - I(V_2^{(2)};S_1|Q=2)\big]\\ &\overset{(b)}{=} I(V_2^{(Q)};S_2|Q) - I(V_2^{(Q)};S_1|Q)\\ &\overset{(c)}{=} I(V_2^{(Q)},Q;S_2) - I(V_2^{(Q)},Q;S_1), \end{aligned} \tag{110}$$

and

$$\begin{aligned}\alpha C_2^{lb}(R'^{(1)}) + \bar{\alpha}C_2^{lb}(R'^{(2)}) =&\alpha\big[I(U^{(1)};Y^{(1)},S_2|V_2^{(1)}) - I(U^{(1)};S_1|V_2^{(1)})\big]\\ &+ \bar{\alpha}\big[I(U^{(2)};Y^{(2)},S_2|V_2^{(2)}) - I(U^{(2)};S_1|V_2^{(2)})\big]\\ \overset{(d)}{=}&I(U^{(Q)};Y^{(Q)},S_2|V_2^{(Q)},Q) - I(U^{(Q)};S_1|V_2^{(Q)},Q),\end{aligned} \tag{111}$$

where $(a),(b),(c)$ and (d) all follow from the fact that Q is independent of (S_1,S_2,V_2,U,X,Y) and from Q's probability distribution. Now, let $V_2' = (V_2^{(Q)},Q), U' = U^{(Q)}, Y' = Y^{(Q)}$ and $X' = X^{(Q)}$. Then, following from the equalities above, for any two rates $R'^{(1)}$ and $R'^{(2)}$ and for any $0 \leq \alpha \leq 1$, there exists a set of random variables (U',V_2',X',Y') that maintains

$$\alpha R'^{(1)} + \bar{\alpha}R'^{(2)} = I(V_2';S_2) - I(V_2';S_1), \tag{112}$$

and

$$\begin{aligned}C_2^{lb}\big(\alpha R'^{(1)} + \bar{\alpha}R'^{(2)}\big) \geq& I(U';Y',S_2|V_2') - I(U';S_1|V_2')\\ =&\alpha C_2^{lb}(R'^{(1)}) + \bar{\alpha}C_2^{lb}(R'^{(2)}).\end{aligned} \tag{113}$$

This completes the proof of the concavity of $C_2^{lb}(R')$ in R'. □

Part 2: We prove here that it is enough to take X to be a deterministic function of (U, S_1, V_1) in order to maximize $I(U;Y,S_2,V_1) - I(U;S_1,V_1)$. Fix $p(u, v_1|s_1)$. Note that

$$\begin{aligned} p(y, s_2|u, v_1) &= \sum_{x,s_1} p(s_1|, u, v_1)p(s_2|s_1, v_1, u)p(x|s_1, s_2, v_1, u)p(y|x, s_1, s_2, v_1, u) \\ &= \sum_{x,s_1} p(s_1|u, v_1)p(s_2|s_1)p(x|s_1, v_1, u)p(y|x, s_1, s_2) \end{aligned} \tag{114}$$

is linear in $p(x|u, v_1, s_1)$. This follows from the fact that fixing $p(u, v_1|s_1)$ also defines $p(s_1|u, v_1)$ and from the following Markov chains $S_2 - S_1 - (V_1, U)$, $X - (S_1, V_1, U) - S_2$ and $Y - (X, S_1, S_2) - (V_1, U)$. Hence, since $I(U;Y,S_2|V_1)$ is convex in $p(y, s_2|v_1)$ it is also convex in $p(x|u, v_1, s_1)$. Noting also that $I(U;S_1|V_1)$ is constant given a fixed $p(u, v_1|s_1)$, we can conclude that $I(U;Y,S_2|V_1) - I(U;S_1|V_1)$ is convex in $p(x|u, v_1, s_1)$ and, hence, it gets its maximum at the boundaries of $p(x|u, v_1, s_1)$, i.e., when the last is equal 0 or 1. This implies that X can be expressed as a deterministic function of (U, V_1, S_1). □

Part 3: We prove now the cardinality bound for Theorem 1. First, let us recall the support lemma [31, p.310]. Let $\mathcal{P}(\mathcal{Z})$ be the set of PMFs on the set $\mathcal{Z}$, and let the set $\mathcal{P}(\mathcal{Z}|\mathcal{Q}) \subseteq \mathcal{P}(\mathcal{Z})$ be a collection of PMFs $p(z|q)$ on $\mathcal{Z}$ indexed by $q \in \mathcal{Q}$. Let $g_j,\ j = 1, \ldots, k$, be continuous functions on $\mathcal{P}(\mathcal{Z}|\mathcal{Q})$. Then, for any $Q \sim F_Q(q)$, there exists a finite random variable $Q' \sim p(q')$ taking at most k values in $\mathcal{Q}$ such that

$$\begin{aligned} \mathbb{E}\Big[g_j(p_{Z|Q}(z|Q))\Big] &= \int_{\mathcal{Q}} g_j(p_{Z|Q}(z|q))\mathrm{d}F(q) \\ &= \sum_{q'} g_j(p_{Z|q}(z|q'))p(q'). \end{aligned} \tag{115}$$

We first reduce the alphabet size of V_1 while considering the alphabet size of U to be constant and then we calculate the cardinality of U. Consider the following continuous functions of $p(x, s_1, s_2, u|v_1)$

$$g_j = \begin{cases} P_{XS_1S_2|V}(j|v_1), & j \in \big\{1, 2, \ldots, |\mathcal{X}||\mathcal{S}_1||\mathcal{S}_2| - 1\big\}, \\ I(V_1;S_1) - I(V_1;Y,S_2) & j = |\mathcal{X}||\mathcal{S}_1||\mathcal{S}_2|, \\ I(U;Y,S_2|V_1 = v_1) - I(U;S_1|V_1 = v_1) & j = |\mathcal{X}||\mathcal{S}_1||\mathcal{S}_2| + 1. \end{cases} \tag{116}$$

Then, by the support lemma, there exists a random variable V_1' with $|\mathcal{V}_1'| \leq |\mathcal{X}||\mathcal{S}_1||\mathcal{S}_2| + 1$ such that $p(x, s_1, s_2)$, $I(V_1;S_1) - I(V_1;Y,S_2)$ and $I(U;Y,S_2|V_1) - I(U;S_1|V_1)$ are preserved. Notice that the probability of U might have changed due to changing V_1; we denote the corresponding U as U'. Next, for $v_1' \in \mathcal{V}_1'$ and the corresponding probability $p(v_1')$ that we found in the previous step, we consider $|\mathcal{X}||\mathcal{S}_1||\mathcal{S}_2||\mathcal{V}_1'|$ continuous functions of $p(x, s_1, s_2, v_1'|u')$

$$f_j = \begin{cases} P_{XS_1S_2V_1'|U'}(j|u') & j = \big\{1, 2, \ldots, |\mathcal{X}||\mathcal{S}_1||\mathcal{S}_2||\mathcal{V}_1'| - 1\big\}, \\ I(U';Y,S_2|V_1') - I(U';S_1|V_1') & j = |\mathcal{X}||\mathcal{S}_1||\mathcal{S}_2||\mathcal{V}_1'|. \end{cases} \tag{117}$$

Thus, there exists a random variable U'' with $|\mathcal{U}''| \leq |\mathcal{X}||\mathcal{S}_1||\mathcal{S}_2||\mathcal{V}_1'|$ such that the mutual information expressions above and all the desired Markov conditions are preserved. Notice that the expression $I(V_1;S_1) - I(V_1;Y,S_2)$ is being preserved since $p(x, s_1, s_2, v_1')$ is being preserved.

To conclude, we can bound the cardinality of the auxiliary random variables of Theorem 1 Case 1 by $|\mathcal{V}_1| \leq |\mathcal{X}||\mathcal{S}_1||\mathcal{S}_2|+1$ and $|\mathcal{U}| \leq |\mathcal{X}||\mathcal{S}_1||\mathcal{S}_2||\mathcal{V}_1| \leq |\mathcal{X}||\mathcal{S}_1||\mathcal{S}_2|\big(|\mathcal{X}||\mathcal{S}_1||\mathcal{S}_2|+1\big)$ without limiting the generality of the solution. □

APPENDIX E
PROOF OF THEOREM 3

Proof: First, let us formulate the Lagrangian for the primal optimization problem defined in (40):

$$\begin{aligned} L\big(\mathbf{q},\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}\big) = &\sum_{x,s,t} p(x,s)q(t|x)\log\frac{q(t|x)}{Q(t|s)} \\ &+\sum_{x}\mu_x\Big(\sum_{t} q(t|x)-1\Big) \\ &+\gamma\Big(\sum_{x,s,t} p(x,s)q(t|x)d\big(x,t(s)\big)-D\Big) \\ &-\sum_{x,t}\lambda_{x,t}q(t|x), \end{aligned} \tag{118}$$

with Lagrange multipliers $\boldsymbol{\mu},\gamma \geq 0$ and $\boldsymbol{\lambda} \succeq 0$. Recall that $Q(t|s)$ is a marginal distribution that corresponds with $q(t|x)$. i.e.,

$$Q(t|s) = \frac{\sum_x p(x,s)q(t|x)}{\sum_s p(x,s)}. \tag{119}$$

In addition, recall the definition of the Lagrange dual function,

$$g\big(\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}\big) = \inf_{\mathbf{q}} L\big(\mathbf{q},\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}\big). \tag{120}$$

In the following proof, we use $\mathbf{q}^*_{\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}}$ to denote the optimal minimizer of the Lagrangian, $L\big(\mathbf{q},\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}\big)$, for any fixed $\boldsymbol{\mu},\gamma$, and $\boldsymbol{\lambda}$. We also use the notation $g\big(\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}\big|\mathbf{q}^*_{\mu,\gamma,\lambda}\big)$ to denote the Lagrange dual function with $\mathbf{q}^*_{\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}}$ as a constant parameter.

The outline of the proof is as follows: we first find the PMF $\mathbf{q}^*_{\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}}$, which is the minimizer of the Lagrangian, $L\big(\mathbf{q},\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}\big)$. We then formulate the Lagrange dual function, $g\big(\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}\big|\mathbf{q}^*_{\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}}\big)$, and the Lagrange dual problem, which is to maximize g over $\boldsymbol{\mu},\gamma \geq 0$ and $\boldsymbol{\lambda} \succeq 0$. Next, we argue that we can maximize g over $\boldsymbol{\mu},\gamma \geq 0, \boldsymbol{\lambda} \succeq 0$ and, in addition, over any $\mathbf{q}$ that nullifies the derivative of the Lagrangian (i.e., maintains equation (123)) without increasing the solution of the Lagrange dual problem. We then note that it is possible to write the Lagrange dual problem with the variable $p(x|s,t)$ instead of $q(t|x)$, where $p(x|s,t)$ is a marginal distribution associated with $q(t|x)$. i.e., $p(x|s,t) = \frac{p(x,s)q(t|x)}{\sum_{s,t} p(x,s)q(t|x)}$ is constrained to maintains the Markov chain $T-X-S$. Our next key step is to prove that we can omit the Markov chain constraint without increasing the maximal value of the Lagrange dual problem. We then conclude our proof by formulating the Lagrange dual problem that we obtained in a geometric programming convex form.

In order to formulate $g\big(\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}\big)$, we first find the PMF $\mathbf{q}^*_{\mu,\gamma,\lambda}$ that minimizes the Lagrangian, $L\big(\mathbf{q},\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}\big)$, which is a convex function of $\mathbf{q}$. First, notice that

$$\begin{aligned}
&\frac{\partial}{\partial q(t|x)} \sum_{x',s',t'} p(x',s')q(t'|x') \log \frac{q(t'|x')}{Q(t'|s')} \\
&\quad \overset{(a)}{=} \sum_{s'} p(x,s') \log \frac{q(t|x)}{Q(t|s')} + \sum_{s'} p(x,s') - \sum_{x',s'} p(x',s')q(t|x') \frac{p(x,s')}{p(s')} \frac{1}{Q(t|s')} \\
&\quad = \sum_{s'} p(x,s') \log \frac{q(t|x)}{Q(t|s')} + p(x) - \sum_{s'} p(x,s') \sum_{x'} p(x',s')q(t|x') \frac{1}{p(s')} \frac{1}{Q(t|s')} \\
&\quad \overset{(b)}{=} \sum_{s'} p(x,s') \log \frac{q(t|x)}{Q(t|s')} + p(x) - \sum_{s'} p(x,s') \\
&\quad = \sum_{s'} p(x,s') \log \frac{q(t|x)}{Q(t|s')},
\end{aligned} \tag{121}$$

where (a) follows from the fact that

$$\begin{aligned}
\frac{\partial Q(t'|s')}{\partial q(t|x)} &= \frac{\partial}{\partial q(t|x)} \frac{\sum_{x''} p(x'',s')q(t'|x'')}{p(s')} \\
&= \begin{cases} \frac{p(x,s')}{p(s')}, & t' = t \\ 0, & t' \neq t \end{cases},
\end{aligned} \tag{122}$$

and (b) follows from the fact that $p(x,s')$ is independent of x' and the fact that $\sum_{x'} p(x',s')q(t|x')\frac{1}{p(s')} = Q(t|s')$.

Next, we formulate the derivative of the Lagrangian with respect to $q(t|x)$ and we constrain it to be equal to 0.

$$\frac{\partial L}{\partial q(t|x)} = \sum_s p(x,s) \log \frac{q(t|x)}{Q(t|s)} + \mu_x + \gamma \sum_s p(x,s) d\big(x,t(s)\big) - \lambda_{x,t} = 0. \tag{123}$$

Using elementary mathematical manipulations we get

$$\log q(t|x) = \sum_s p(s|x) \Big[\log Q(t|s) - \frac{\mu_x}{p(x)} - \gamma d\big(x,t(x)\big) - \frac{\lambda_{x,t}}{p(x)} \Big]. \tag{124}$$

Hence,

$$q^*_{\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}}(t|x) = \prod_s \Big[Q^*_{\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}}(t|s) \exp\Big\{ -\frac{\mu_x}{p(x)} - \gamma d\big(x,t(s)\big) + \frac{\lambda_{x,t}}{p(x)} \Big\} \Big]^{p(s|x)} \tag{125}$$

is an optimal minimizer of the Lagrangian. We get the Lagrange dual function by substituting $\mathbf{q}$ in the Lagrangian with $\mathbf{q}^*_{\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}}$ that we got in (125) and by using constraint (123).

$$\begin{aligned}
g\big(\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}\big|\mathbf{q}^*_{\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}}\big) &= \inf_{\mathbf{q}} L\big(\mathbf{q},\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}\big) \\
&= L\big(\mathbf{q}^*_{\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}},\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}\big) \\
&= \begin{cases} -\sum_x \mu_x - \gamma D, & \sum_s p(x,s) \log \frac{q^*_{\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}}(t|x)}{Q^*_{\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}}(t|s)} + \mu_x + \gamma \sum_s p(x,s) d\big(x,t(s)\big) - \lambda_{x,t} = 0 \\ & \qquad \forall x,t \\ -\infty, & \text{otherwhise} \end{cases}
\end{aligned} \tag{126}$$

We get the Lagrange dual problem by making the constraints explicit:

$$\begin{array}{ll} \text{maximize} & -\sum_x \mu_x - \gamma D \\ \text{subject to} & \sum_s p(x,s)\log\frac{q^*_{\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}}(t|x)}{Q^*_{\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}}(t|s)} + \mu_x + \gamma\sum_s p(x,s)d\big(x,t(s)\big) - \lambda_{x,t} = 0,\ \forall x,t, \\ & \gamma \geq 0, \\ & \lambda_{x,t} \geq 0,\ \forall x,t, \end{array} \tag{127}$$

where the maximization variables are $\boldsymbol{\mu},\gamma$ and $\boldsymbol{\lambda}$ and the constant parameters are the PMFs $\mathbf{q}^*_{\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}}$ and $p(x,s)$, the distortion measure $d\big(x,t(s)\big)$ and the distortion constraint D. Notice that since the primal problem, (40), is a convex problem with an optimal value of $R(D)$, then the solution of (127) is a lower bound on $R(D)$ [28, Chapter 5.2.2], and, if Slater's condition holds, then strong duality holds and the optimal value of (127) is $R(D)$.

Now, notice that any $\mathbf{q}$ that maintains the first inequality constraint in (127) nullifies the derivative of the Lagrangian and, hence, results in the same value when placed in the Lagrangian; this value is exactly the Lagrange dual function. Therefore, since g gets the same value for any $\mathbf{q}$ that maintains the constraint (123), we can maximize g over all PMFs $\mathbf{q}$ that maintain constraint (123) without changing g's value. Consequently, the Lagrange dual problem in (127) becomes:

$$\begin{array}{ll} \text{maximize} & -\sum_x \mu_x - \gamma D \\ \text{subject to} & \sum_s p(x,s)\log\frac{q(t|x)}{Q(t|s)} + \mu_x + \gamma\sum_s p(x,s)d\big(x,t(s)\big) - \lambda_{x,t} = 0,\ \forall x,t, \\ & \gamma \geq 0, \\ & \lambda_{x,t} \geq 0,\ \forall x,t, \\ & \sum_t q(t|x) = 1,\ \forall x, \end{array} \tag{128}$$

where the maximization variables are $\boldsymbol{\mu},\gamma,\boldsymbol{\lambda}$ and $\mathbf{q}$ and the constant parameters are $p(x,s)$, $d\big(x,t(s)\big)$ and D.

Next, combining (125) and the fact that $Q(t|s) \geq 0$, we get that we can replace the first constraint in (128) with

$$q(t|x) = \prod_s \left[Q(t|s)\exp\left\{-\frac{\mu_x}{p(x)} - \gamma d\big(x,t(s)\big) + \frac{\lambda_{x,t}}{p(x)}\right\}\right]^{p(s|x)},\ \forall x,t. \tag{129}$$

Since $q(t|x)$ is independent of s, we can state that

$$1 = \prod_s \left[\frac{Q(t|s)}{q(t|x)}\exp\left\{-\frac{\mu_x}{p(x)} - \gamma d\big(x,t(s)\big) + \frac{\lambda_{x,t}}{p(x)}\right\}\right]^{p(s|x)}. \tag{130}$$

Let us denote $\alpha_x = -\frac{\mu_x}{p(x)}$ and note that $\frac{Q(t|s)}{q(t|x)} = \frac{p(x|s)Q(t|s)}{p(t,x|s)} = \frac{p(x|s)}{p(x|s,t)}$, where $p(x|s,t)$ maintains the Markov chain $T-X-S$. Therefore, equation (130) becomes

$$1 = \prod_s \left[p(x|s)\exp\left\{\alpha_x - \gamma d\big(x,t(s)\big) + \frac{\lambda_{x,t}}{p(x)} - \log p(x|s,t)\right\}\right]^{p(s|x)}, \tag{131}$$

for all x, t, and the Lagrange dual problem can be reformulated as

$$\begin{array}{ll} \text{maximize} & \sum_x \alpha_x p(x) - \gamma D \\ \text{subject to} & 1 = \prod_s \left[p(x|s) \exp\left\{ \alpha_x - \gamma d\big(x, t(s)\big) + \frac{\lambda_{x,t}}{p(x)} - \log p(x|s,t) \right\} \right]^{p(s|x)}, \ \forall x, t \\ & \gamma \geq 0, \\ & \sum_t p(x|s,t) = 1, \ \forall x, \\ & p(x|s,t) \text{ maintain the Markov chain } T - X - S, \end{array} \tag{132}$$

where the variables of the maximization are $\boldsymbol{\alpha}, \gamma, \boldsymbol{\lambda}$ and $\mathbf{p} \in \mathbb{R}^{|\mathcal{X}||\mathcal{S}||\mathcal{T}|}$, which is the set of all $p(x|s,t)$ for all $x \in \mathcal{X}, s \in \mathcal{S}$ and $t \in \mathcal{T}$, and the constant variables are $p(x,s)$, $d\big(x, t(s)\big)$ and D. Notice that (132) is not a convex problem anymore, since the constraint functions are not convex. We deal with this problem in the following steps by using geometric programming principles.

Next, we want to prove that it is possible to maximize (132) over any PMF, $\mathbf{p}$. i.e., we want to prove that dropping the last constraint in (132) does not change the validity of the solution.

First, since (132) is an equivalent Lagrange dual problem, then, according to [28, Chapter 5.2.2], we can state that for any choice of $\boldsymbol{\alpha}, \gamma$ and $\boldsymbol{\lambda}$ it yields a lower bound on $R(D)$. Furthermore, according to [28, Chapter 5.2.3], if Slater's condition holds, then the solution of (132) coincides with $R(D)$, which is the optimal solution of the primal problem. Now, dropping the constraint that the Markov chain $T - X - S$ must hold, necessarily allows the optimal solution of (132) to be greater than or equal to the solution where $T - X - S$ holds. We are left to prove that maximizing over any PMF, $\mathbf{p}$, cannot exceed $R(D)$. Let us place $p(x|s,t) = \frac{p(t|x,s)p(x|s)}{p(t|s)}$ in (131) and look at the following inequalities:

$$\begin{aligned} 1 &= \prod_s \left[p(x|s) \exp\left\{ \alpha_x - \gamma d\big(x, t(s)\big) + \frac{\lambda_{x,t}}{p(x)} - \log \frac{p(t|x,s)p(x|s)}{p(t|s)} \right\} \right]^{p(s|x)} \\ &= \prod_s \left[\exp\left\{ \log p(x|s) + \alpha_x - \gamma d\big(x, t(s)\big) + \frac{\lambda_{x,t}}{p(x)} - \log \frac{p(t|x,s)p(x|s)}{p(t|s)} \right\} \right]^{p(s|x)} \\ &= \exp\left\{ \alpha_x - \gamma \sum_s p(s|x) d\big(x, t(s)\big) + \frac{\lambda_{x,t}}{p(x)} - \sum_s p(s|x) \log p(t|x,s) + \sum_s p(s|x) \log p(t|s) \right\} \\ &\overset{(a)}{\geq} \exp\left\{ \alpha_x - \gamma \sum_s p(s|x) d\big(x, t(s)\big) + \frac{\lambda_{x,t}}{p(x)} - \log \Big(\sum_s p(s|x) p(t|x,s) \Big) + \sum_s p(s|x) \log p(t|s) \right\} \\ &= \exp\left\{ \alpha_x - \gamma \sum_s p(s|x) d\big(x, t(s)\big) + \frac{\lambda_{x,t}}{p(x)} - \log p(t|x) + \sum_s p(s|x) \log p(t|s) \right\} \\ &\overset{(b)}{=} \exp\left\{ \alpha_x - \gamma \sum_s p(s|x) d\big(x, t(s)\big) + \frac{\lambda_{x,t}}{p(x)} - \sum_s p(s|x) \log \frac{p(t|x)p(x|s)}{p(t|s)} + \sum_s p(s|x) \log p(x|s) \right\} \\ &= \prod_s \left[p(x|s) \exp\left\{ \alpha_x - \gamma d\big(x, t(s)\big) + \frac{\lambda_{x,t}}{p(x)} - \log \frac{p(t|x)p(x|s)}{p(t|s)} \right\} \right]^{p(s|x)}, \end{aligned} \tag{133}$$

where (a) follows from Jensen's inequality and (b) follows from the fact that $p(t|x)$ is independent of s. Notice that by reducing the value of $\sum_s p(s|x) \log p(t|x,s)$, we allow $\alpha_x - \gamma \sum_s p(s|x) d\big(x, t(s)\big)$ to be greater and, hence, we improve our maximum. Therefore, for any $p(x|s,t) = \frac{p(t|s)}{p(t|x,s)p(x|s)}$, we can take $p'(x|s,t) = \frac{p(t|s)}{p(x|s) \sum_{s'} p(s'|x) p(t|x,s')}$, which satisfies the Markov chain $T-X-S$, and that the maximum over $p(t|x)=\sum_s p(s|x)p(t|x,s)$ would be equal to or greater than the maximum over $p(x|s,t)$. This, and the fact that maximizing over $p(t|x)$ cannot exceed $R(D)$ and that $R(D)$ can be achieved by using $p^*(x|s,t)$ that corresponds to $q^*(t|x)$, prove that, indeed, we can maximize over $p(x|s,t)$ without changing the result of the maximization. Therefore, our dual problem now becomes

$$\begin{array}{ll} \text{maximize} & \sum_x \alpha_x p(x) - \gamma D \\ \text{subject to} & \prod_s \left[p(x|s) \exp\left\{ \alpha_x - \gamma d\big(x,t(s)\big) + \frac{\lambda_{x,t}}{p(x)} - \log p(x|s,t) \right\} \right]^{p(s|x)} = 1 \quad \forall x,t, \\ & \sum_x p(x|s,t) = 1 \quad \forall s,t \\ & \gamma \geq 0. \end{array} \tag{134}$$

In order to make the problem convex, we need to convert the equality constraints that are not affine into inequality constraints. Let us go back to (131); since $\lambda_{x,t} \geq 0$ for all x and t and since $p(x,s) \geq 0$, the constraint (131) can be replaced by

$$1 \geq \prod_s \left[p(x|s) \exp\left\{ \alpha_x - \gamma d\big(x,t(s)\big) - \log p(x|s,t) \right\} \right]^{p(s|x)} \tag{135}$$

without changing the solution of (132). Next, notice that there is a tradeoff between $-\log p(x|s,t)$ and $\alpha_x - \gamma d\big(x,t(s)\big)$. Therefore, we expect $-\log p(x|s,t)$ to be as small as possible to allow $\alpha_x - \gamma d\big(x,t(s)\big)$ to be as large as possible. Hence, we can replace the constraint

$$\sum_x p(x|s,t) = 1 \quad \forall s,t, \tag{136}$$

which is equivalent to

$$\sum_x \exp\big\{ \log p(x|s,t) \big\} = 1 \quad \forall s,t, \tag{137}$$

with the weaker constraint

$$\sum_x \exp\big\{ \log p(x|s,t) \big\} \leq 1 \quad \forall s,t, \tag{138}$$

without changing the result of the maximization. We denote $y_{x,t,s} = \log p(x|s,t)$ and rewrite the dual problem as

$$\begin{array}{ll} \text{maximize} & \sum_x \alpha_x p(x) - \gamma D \\ \text{subject to} & \prod_s \left[p(x|s) \exp\left\{ \alpha_x - \gamma d\big(x,t(s)\big) - y_{x,s,t} \right\} \right]^{p(s|x)} \leq 1 \quad \forall x,t, \\ & \sum_x \exp\left\{ y_{x,s,t} \right\} \leq 1 \quad \forall s,t, \\ & \gamma \geq 0, \end{array} \tag{139}$$

where the variables of the maximization are $\boldsymbol{\alpha}, \gamma$ and $\mathbf{y}$ and the constant parameters are the PMF, $p(x,s)$, the distortion measure, $d\big(x,t(s)\big)$, and the distortion constraint, D.

Lastly, we present the dual problem in a geometric programming convex form by taking $\log(\cdot)$ on the first two inequality constraints:

$$\begin{aligned} &\text{maximize} && \textstyle\sum_x \alpha_x p(x) - \gamma D \\ &\text{subject to} && \alpha_x + \textstyle\sum_s p(s|x)\Big[\log p(x|s) - \gamma d\big(x,t(s)\big) - y_{x,s,t}\Big] \le 0 \quad \forall x,t, \\ & && \log\big(\textstyle\sum_x \exp\{y_{x,s,t}\}\big) \le 0 \quad \forall s,t, \\ & && \gamma \ge 0, \end{aligned} \tag{140}$$

where the variables of the maximization are $\boldsymbol{\alpha}, \gamma$ and $\mathbf{y}$ and the constant parameters are $p(x,s), d\big(x,t(s)\big)$ and D.

■

APPENDIX F
PROOFS FOR SECTION VI

A. Proof of Lemma 4

Proof: For $0 \le \alpha \le 1$ and $\bar{\alpha} = 1-\alpha$

$$\begin{aligned} J_w(\alpha q_1 + \bar{\alpha} q_2, \alpha Q_1 + \bar{\alpha} Q_2) &= \sum_{s_1,s_2,v_2,t,y} p(s_1,s_2)w(v_2|s_2)p(y|t,s_1,s_2,v_2)\Big(\alpha q_1 + \bar{\alpha} q_2\Big)\log\frac{\alpha Q_1 + \bar{\alpha} Q_2}{\alpha q_1 + \bar{\alpha} q_2} \\ &\overset{(a)}{\le} \sum_{s_1,s_2,v_2,t,y} p(s_1,s_2)w(v_2|s_2)p(y|t,s_1,s_2,v_2)\Big(\alpha q_1 \log\frac{Q_1}{q_1} + \bar{\alpha} q_2 \log\frac{Q_2}{q_2}\Big) \\ &= \alpha J_w(q_1,Q_1) + \bar{\alpha} J_w(q_2,Q_2), \end{aligned} \tag{141}$$

where (a) follows from the log-sum inequality:

$$\sum_i a_i \log\frac{a_i}{b_i} \ge a\log\frac{a}{b}, \tag{142}$$

for $\sum_i a_i = a$ and $\sum_i b_i = b$. ■

B. Proof of Lemma 6

Proof: Let us calculate q^* using the KKT conditions. We want to maximize $J_w(q^*,Q)$ over q^*, where for all t, s_1 and v_2, $0 \le q^*(t|s_1,v_2) \le 1$ and $\sum_{t'} q^*(t'|s_1,v_2) = 1$.

For fixed s_1 and v_2,

$$0 = \frac{\partial}{\partial q^*}\Big(J_w(q^*,Q) + \big(1 - \sum_t q^*(t|s_1,v_2)\big)\nu_{s_1,v_2}\Big) \tag{143}$$

$$= \sum_{s_2,y} p(s_1,s_2)w(v_2|s_2)p(y|t,s_1,s_2,v_2)\Big(\log\frac{Q(t|y,s_2,v_2)}{q^*(t|s_1,v_2)} - 1\Big) - \nu_{s_1,v_2}, \tag{144}$$

divide by $p(s_1,v_2)$,

$$0 = -\log q^*(t|s_1,v_2) + \frac{\sum_{s_2,y} p(s_1,s_2)w(v_2|s_2)p(y|t,s_1,s_2,v_2)}{p(s_1,v_2)}\log Q(t|y,s_2,v_2) - 1 + \frac{\nu_{s_1 v_2}}{p(s_1,v_2)}, \tag{145}$$

define $-1+\frac{\nu_{s_1v_2}}{p(s_1,v_2)}=\log\nu'_{s_1,v_2}$, hence

$$q^*(t|s_1,v_2)=\nu'_{s_1,v_2}\prod_{s_2,y}Q(t|y,s_2,v_2)^{p(s_2|s_1,v_2)p(y|t,s_1,s_2,v_2)}, \tag{146}$$

and from the constraint $\sum_{t'}q^*(t'|s_1,v_2)=1$ we get that

$$q^*(t|s_1,v_2)=\frac{\prod_{s_2,y}Q(t|y,s_2,v_2)^{p(s_2|s_1,v_2)p(y|t,s_1,s_2,v_2)}}{\sum_{t'}\prod_{s_2,y}Q(t'|y,s_2,v_2)^{p(s_2|s_1,v_2)p(y|t',s_1,s_2,v_2)}}. \tag{147}$$

■

C. *Proof of Lemma 7*

The proof for this lemma is done in three steps: first, we prove that $U_w(q_1)$ is greater than or equal to $J_w(q_0,Q_0^*)$ for any two PMFs $q_0(t|s_1,v_2)$ and $q_1(t|s_1,v_2)$, then, we use Lemma 3 and Lemma 5 to state that for the optimal PMF, $q_c(t|s_1,v_2)$, $C_{2,w}^{lb}=J_w(q_c,Q_c^*)$, and, therefore, $U_w(q)$ is an upper bound of $C_{2,w}^{lb}$ for every $q(t|s_1,v_2)$. Thirdly, we prove that $U_w(q)$ converges to $C_{2,w}^{lb}$.

Proof: Consider any two PMFs, $q_0(t|s_1,v_2)$ and $q_1(t|s_1,v_2)$, their corresponding $\{p_0(s_1,s_2,v_2,t,y),Q_0^*(t|y,s_2,v_2)\}$ and $\{p_1(s_1,s_2,v_2,t,y),Q_1^*(t|y,s_2,v_2)\}$, respectively, according to (50) and (52) and consider also the following inequalities:

$$\begin{aligned}
&\sum_{s_1,s_2,v_2,t,y}p_0(s_1,s_2,v_2,t,y)\log\frac{Q_1^*(t|y,s_2,v_2)}{q_1(t|s_1,v_2)}-J_w(q_0,Q_0^*)\\
&=\sum_{s_1,s_2,v_2,t,y}p_0(s_1,s_2,v_2,t,y)\Big(\log\frac{Q_1^*(t|y,s_2,v_2)}{q_1(t|s_1,v_2)}-\log\frac{Q_0^*(t|y,s_2,v_2)}{q_0(t|s_1,v_2)}\Big)\\
&=\sum_{s_1,s_2,v_2,t,y}p_0(s_1,s_2,v_2,t,y)\log\Big(\frac{Q_1^*(t|y,s_2,v_2)}{Q_0^*(t|y,s_2,v_2)}\frac{q_0(t|s_1,v_2)}{q_1(t|s_1,v_2)}\Big)\\
&=\mathbb{D}\big(q_0(t|s_1,v_2)\big\|q_1(t|s_1,v_2)\big)-\mathbb{D}\big(Q_0^*(t|y,s_2,v_2)\big\|Q_1^*(t|y,s_2,v_2)\big)\\
&\overset{(a)}{=}\mathbb{D}\big(q_0(t|s_1,s_2,v_2)p(y|t,s_1,s_2,v_2)p(s_1,s_2)w(v_2|s_2)\big\|q_1(t|s_1,s_2,v_2)p(y|t,s_1,s_2,v_2)p(s_1,s_2)w(v_2|s_2)\big)\\
&\qquad-\mathbb{D}\big(Q_0^*(t|y,s_2,v_2)\big\|Q_1^*(t|y,s_2,v_2)\big)\\
&=\mathbb{D}\big(p_0(s_1,s_2,v_2,t,y)\big\|p_1(s_1,s_2,v_2,t,y)\big)-\mathbb{D}\big(Q_0^*(t|y,s_2,v_2)\big\|Q_1^*(t|y,s_2,v_2)\big)\\
&\overset{(b)}{=}\mathbb{D}\big(p_0(s_2,v_2,y)Q_0^*(t|y,s_2,v_2)p_0(s_1|s_2,v_2,t,y)\big\|p_1(s_2,v_2,y)Q_1^*(t|y,s_2,v_2)p_1(s_1|s_2,v_2,t,y)\big)\\
&\qquad-\mathbb{D}\big(Q_0^*(t|y,s_2,v_2)\big\|Q_1^*(t|y,s_2,v_2)\big)\\
&=\mathbb{D}\big(p_0(s_2,v_2,y)\big\|p_1(s_2,v_2,y)\big)+\mathbb{D}\big(p_0(s_1|s_2,v_2,t,y)\big\|p_1(s_1|s_2,v_2,t,y)\big)\\
&\overset{(c)}{=}\geq 0,
\end{aligned} \tag{148}$$

where $\mathbb{D}(\cdot\|\cdot)$ is the K-L divergence, $p_j(s_2,v_2,y)$ and $p_j(s_1|s_2,v_2,t,y)$ are marginal distributions of $p_j(s_1,s_2,v_2,t,y)$ for $j=0,1$, (a) follows from the fact that T is independent of S_2 given (S_1,V_2) and from the K-L divergence properties, (b) follows from the fact that $Q_j^*(t|y,s_2,v_2)$ is a marginal distribution of $p_j(s_1,s_2,v_2,t,y)$ for $j=0,1$ and (c) follows from the fact that $\mathbb{D}(\cdot\|\cdot)\geq 0$ always.

Thus,

$$\begin{aligned}
J(q_0, Q_0^*) \leq &\sum_{s_1,s_2,v_2,t,y} p_0(s_1, s_2, v_2, t, y) \log \frac{Q_1^*(t|y, s_2, v_2)}{q_1(t|s_1, v_2)} \\
= &\sum_{s_1,s_2,v_2,t,y} p(s_1, s_2)w(v_2|s_2)q_0(t|s_1, v_2)p(y|t, s_1, s_2, v_2) \log \frac{Q_1^*(t|y, s_2, v_2)}{q_1(t|s_1, v_2)} \\
= &\sum_{s_1,v_2} p(s_1, v_2) \sum_t q_0(t|s_1, v_2) \sum_{s_2} p(s_2|s_1, v_2) \sum_y p(y|t, s_1, s_2, v_2) \log \frac{Q_1^*(t|y, s_2, v_2)}{q_1(t|s_1, v_2)} \\
\leq &\sum_{s_1,v_2} p(s_1, v_2) \max_{t'} \sum_{s_2} p(s_2|s_1, v_2) \sum_y p(y|t', s_1, s_2, v_2) \log \frac{Q_1^*(t'|y, s_2, v_2)}{q_1(t'|s_1, v_2)} \\
= &U_w(q_1). \qquad (149)
\end{aligned}$$

We proved that $U_w(q_1)$ is greater than or equal to $J_w(q_0, Q_0^*)$ for any choice of $q_0(t|s_2, v_2)$ and $q_1(t|s_1, v_2)$. Therefore, by taking $q_0(t|s_1, v_2)$ to be the distribution that achieves $C_{2,w}^{lb}$ and by considering Lemma 3 and Lemma 5, we conclude that $U_w(q) \geq C_{w,2}$ for any choice of $q(t|s_1, v_2)$.

In order to prove that $U_w(q)$ converges to $C_{2,w}^{lb}$ let us rewrite equation (144) as

$$\sum_{s_2,y} p(s_2|s_1, v_2)p(y|t, s_1, s_2, v_2) \log \frac{Q(t|y, s_2, v_2)}{q^*(t|s_1, v_2)} = \nu'_{s_1,v_2}. \qquad (150)$$

We can see that for a fixed Q, the right hand side of the equation is independent of t. Considering also

$$\begin{aligned}
J_w(q, Q) = &\sum_{s_1,s_2,v_2,t,y} p(s_1, s_2)w(v_2|s_2)q(t|s_1, v_2)p(y|t, s_1, s_2, v_2) \log \frac{Q(^t|y, s_2, v_2)}{q(t|s_1, v_2)} \\
\leq &\sum_{s_1,v_2} p(s_1, v_2) \max_{t'} \sum_{s_2} p(s_2|s_1, v_2) \sum_y p(y|t', s_1, s_2, v_2) \log \frac{Q^*(t'|y, s_2, v_2)}{q(t'|s_1, v_2)}, \qquad (151)
\end{aligned}$$

we can conclude that the equation holds when the PMF q is the PMF that achieves $C_{2,w}^{lb}$. ∎

References

[1] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *Information Theory, IEEE Transactions on*, vol. 22, no. 1, pp. 1 – 10, jan 1976.

[2] Y. Steinberg, "Coding for channels with rate-limited side information at the decoder, with applications," *Information Theory, IEEE Transactions on*, vol. 54, no. 9, pp. 4283 –4295, sept. 2008.

[3] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Control Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[4] C. E. Shannon, "Channels with side information at the transmitter," *IBM J. Res. Dev.*, vol. 2, no. 4, pp. 289–293, 1958.

[5] C. Heegard and A. A. E. Gamal, "On the capacity of computer memory with defects," *IEEE Transactions on Information Theory*, vol. 29, no. 5, pp. 731–739, 1983.

[6] T. M. Cover and M. Chiang, "Duality between channel capacity and rate distortion with two-sided state information," *IEEE Trans. Inf. Theor.*, vol. 48, no. 6, pp. 1629–1638, Sep. 2006. [Online]. Available: http://dx.doi.org/10.1109/TIT.2002.1003843

[7] A. Rosenzweig, Y. Steinberg, and S. Shamai, "On channels with partial channel state information at the transmitter," *Information Theory, IEEE Transactions on*, vol. 51, no. 5, pp. 1817 – 1830, may 2005.

[8] Y. Cemal and Y. Steinberg, "Coding problems for channels with partial state information at the transmitter," *Information Theory, IEEE Transactions on*, vol. 53, no. 12, pp. 4521 –4536, dec. 2007.

[9] G. Keshet, Y. Steinberg, and N. Merhav, “Channel coding in the presence of side information,” *Found. Trends Commun. Inf. Theory*, vol. 4, no. 6, pp. 445–586, 2007.

[10] A. H. Kaspi, “Two-way source coding with a fidelity criterion,” *IEEE Transactions on Information Theory*, vol. 31, no. 6, pp. 735–740, 1985.

[11] H. Permuter, Y. Steinberg, and T. Weissman, “Two-way source coding with a helper,” *Information Theory, IEEE Transactions on*, vol. 56, no. 6, pp. 2905 –2919, june 2010.

[12] T. Weissman and A. E. Gamal, “Source coding with limited-look-ahead side information at the decoder,” *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5218–5239, 2006.

[13] T. Weissman and N. Merhav, “On causal source codes with side information,” *IEEE Transactions on Information Theory*, vol. 51, no. 11, pp. 4003–4013, 2005.

[14] C. E. Shannon, “Coding theorems for a discrete source with a fidelity criterion,” vol. 7, part 4, pp. 142–163, Mar. 1959.

[15] S. S. Pradhan, J. Chou, and K. Ramchandran, “Duality between source coding and channel coding and its extension to the side information case,” *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1181–1203, 2003.

[16] R. Zamir, S. Shamai, and U. Erez, “Nested linear/lattice codes for structured multiterminal binning,” *IEEE Trans. Inf. Theor.*, vol. 48, no. 6, pp. 1250–1276, Sep. 2006. [Online]. Available: http://dx.doi.org/10.1109/TIT.2002.1003821

[17] J. Su, J. Eggers, and B. Girod, “Illustration of the duality between channel coding and rate distortion with side information,” in *Signals, Systems and Computers, 2000. Conference Record of the Thirty-Fourth Asilomar Conference on*, vol. 2, 29 2000-nov. 1 2000, pp. 1841 –1845 vol.2.

[18] R. E. Blahut, “Computation of channel capacity and rate-distortion functions,” *IEEE Trans. Inform. Theory*, vol. 18, pp. 460–473, 1972.

[19] S. Arimoto, “An algorithm for computing the capacity of arbitrary discrete memorylesschannels,” *IEEE Trans. Inform. Theory*, vol. 18, pp. 14–20, 1972.

[20] F. M. J. Willems, “Computation of the wyner-ziv rate-distortion function,” Research Report, July 1983.

[21] F. Dupuis, W. Yu, and F. Willems, “Blahut-arimoto algorithms for computing channel capacity and rate-distortion with side information,” in *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on*, june-2 july 2004, p. 179.

[22] S. Cheng, V. Stankovic, and Z. Xiong, “Computing the channel capacity and rate-distortion function with two-sided state information,” *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4418–4425, 2005.

[23] O. Sumszyk and Y. Steinberg, “Information embedding with reversible stegotext,” in *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, 28 2009-july 3 2009, pp. 2728 –2732.

[24] I. Naiss and H. H. Permuter, “Extension of the blahut-arimoto algorithm for maximizing directed information,” *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 204–222, 2013.

[25] M. Chiang, S. Boyd, and A. Overview, “Geometric programming duals of channel capacity and rate distortion,” *IEEE Trans. Inform. Theory*, vol. 50, pp. 245–258, 2004.

[26] I. Naiss and H. H. Permuter, “Computable bounds for rate distortion with feed forward for stationary and ergodic sources,” *IEEE Transactions on Information Theory*, vol. 59, no. 2, pp. 760–781, 2013.

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & sons, 1991.

[28] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[29] R. W. Yeung, *Information Theory and Network Coding*, 1st ed. Springer Publishing Company, Incorporated, 2008.

[30] T. Berger, “Multiterminal source coding,” in *Information Theory Approach to Communications*, G. Longo, Ed. CSIM Course and Lectures, 1978, pp. 171–231.

[31] I. Csiszar and J. Korner, *Information theory : Coding theorems for discrete memoryless systems*. Academic Press ; Akademiai Kiado, New York : Budapest, 1981.